AD 653280

CENTER FOR THE INFORMATION SCIENCES
LEHIGH UNIVERSITY BETHLEHEM, PENNSYLVANIA

EXPERIMENTAL RETRIEVAL SYSTEMS STUDIES

STATEMENT NO. 1
Distribution of This Document is Unlimited

Report No. 3

The research reported here is supported by NSF Grants No. GN-451 and No. GE-2569, and by ONR Contract Nonr-610(08)

April 1967

ARCHIVE COPY

EXPERIMENTAL RETRIEVAL SYSTEMS STUDIES

Report No. 3

Ronald R. Anderson

An Associativity Technique for Automatically Optimizing Retrieval Results

Andrew J. Kasarda

A Syntactically Oriented Natural Language Document Retrieval System with a Browsability Feature

David M. Reed

Phrase Indexing

EXPERIMENTAL RETRIEVAL SYSTEMS STUDIES

The Center for the Information Sciences has developed and maintains an experimental system for the literature of the information sciences. At present the collection contains about 2,500 documents and is used for instruction, reference, research and experimentation.

Documents are indexed manually and a coordinate index system is used with a controlled thesaurus. Posting, up-dating, author listings, and both associative and non-associative searches are performed on the GE-225 computer. On-line access is facilitated through a Datanet-15 and a MOD-33 ASR Teletype in the Center.

In addition, a growing collection of natural language text on tape is maintained for automatic indexing and abstracting studies.

This series of studies reports experimentation and research on this operating system.

Report No. 1. Robert M. Curtice. Magnetic Tape and Disc File Organizations for Retrieval. July 1966

Report No. 2. Systems Manual for the Experimental Literature Collection and Reference Retrieval System. April 1967

AN ASSOCIATIVITY TECHNIQUE FOR AUTOMATICALLY OPTIMIZING RETRIEVAL RESULTS

by

Ronald R. Anderson

Abstract

An experiment is described which evaluates the effectiveness of an associative search technique for automatically optimizing retrieval results. Originally, the design of the experiment called for testing an existing automatic retrieval method by using the associativity formula in the Center for the Information Sciences (CIS) document retrieval system. However, during the early stages of the experiment, it became apparent an adjustment was necessary if the automatic technique was to remain effective. Such a modification was made, resulting in an associative technique which would enable the CIS system to both automatically expand and automatically narrow the number of documents retrieved by an initial search request. That is, to retrieve documents related to a request even though they may not be indexed by the exact terms of the request, and to satisfy the user's depth-of-search requirement by presenting the documents retrieved in the order of their relevance to the request.

This research was supported by the National Science Foundation under Grant No. GE-2569, and by Office of Naval Research Contract Nonr-(710)08

Table of Contents

Page

A. Introduction.....1-1

B. Design of Experiment.....1-2

C. Experimental Results.....1-7

D. Conclusion.....1-16

Appendix: Ranked Document Listings

Table 1: ARTIFICIAL INTELLIGENCE

Table 2: DISSEMINATION + INFORMATION

Table 3: LEARNING + SYSTEMS

A. Introduction

During the past few years, statistical measures of word association have become a common search tool in automatic information retrieval systems. These statistical measures are derived from formulas which attempt to correlate two given index terms on the basis of their use frequencies and their frequencies of co-occurrence in the documents of a given collection.

The basis for using statistical techniques is the assumption that the assignment of two index-terms to a given document may be interpreted as a small piece of probabilistic evidence that the two terms are correlated. That is, the document is assumed to describe a relationship between the topics denoted by the two index terms. Presumably, by accumulating such small pieces of evidence from a large collection of documents, it is possible to arrive at a meaningful over-all measure of association for any given pair of index terms.

The use of these statistical methods is valuable because they consider the connections and relationships among topics which are particular to the given document collection. In addition, they not only state whether an index term is related to another one, but also provide an estimate of how closely related it is.

In the second edition of Centralization and Documentation [1], Arthur D. Little, Inc. states that the capability to determine measures of association among index terms and documents leads to the potential ability to design systems possessing the following highly desirable capabilities, which are not available in many existing coordinate searching systems:

1. A capability for automatically generalizing a user's request to make it more compatible with the vocabulary of the retrieval system.

2. A capability for automatically matching the user's depth-of-search requirement to system parameters, by ranking the documents presented to users in decreasing order of probable relevance.

3. A potential capability for nearly instantaneous interaction between user and searching machine, without the need for a human intermediary.

The Center for the Information Sciences automatic document retrieval system, although making use of an associative search technique, does not fully exploit these capabilities.

The first capability exists in part, but manual intervention is necessary before the generalized request can be used. A procedure has been implemented [2] which presents the user with a list of index terms associated with the documents retrieved by his search request, the number of times each of these terms has been used in indexing the documents in the collection, the number of times these terms index the retrieved documents, and an association coefficient for each term based on the results of the search. Using this information, as well as any familiarity he might have with the contents of the collection and with the general indexing strategy, the user is able to estimate with reasonable success how many documents he would receive if he were to conduct a proposed search. Employing this technique, he is able in many cases to optimize his retrieval results by manually reformulating his original request.

The second capability does not exist at all in the CIS system, although a ranked list of the documents retrieved would be beneficial to the user, especially when his initial search request has produced more documents than he can use.

The third capability is satisfied in the CIS system when the searching is performed in an on-line environment.

The lack of fulfillment of the first two capabilities, together with the success of existing associative methods and the cry for more experiments on associative searching [3] led to the work described in this report. It was felt that by supplementing the existing associative procedures with an effective automatic technique, the full potential offered by the use of associative searching in the CIS system could be more nearly achieved. The user would then have the option of attempting to optimize his retrieval results through either semi-automatic or fully automatic means.

B. Design of Experiment

In performing the experiment, the CIS associative search program was utilized. For each associative search, this program produces a

term profile, a list of index terms associated with the search term. The association between an index term and the search term is calculated as follows:

$$\text{ASSOCIATIVITY COEFFICIENT} = \frac{f(ab)^2}{f(a) \times f(b)}, \text{ where}$$

f(a) = total occurrence of the profile term in the document collection

f(b) = total occurrence of the search term in the document collection

f(ab) = co-occurrence of the profile term with the search term.

The term profile for the search term "ARTIFICIAL INTELLIGENCE" is given in Figure 1.

When the search is composed of two or more terms linked by the logical Boolean operations of disjunction (v), conjunction (+), and negation (-), the associativity coefficient is still calculated as though the search was composed of one term. That is, f(b) is equal to the number of documents retrieved by the search and f(ab) is the co-occurrence of the profile term in those documents. The term profile for the search "DISSEMINATION + INFORMATION" is given in Figure 2. Eight documents are retrieved by this search statement.

TOT. OCC.	CO-OCC.	TERM	ASSOC. COEFF.
2	1	adaptive	0.0208
3	1	ALGOL	0.0139
12	2	answers	0.0139
24	24	artificial intelligence	1.0000
76	1	associat(ion,ive)	0.0005
14	2	automata	0.0119
116	1	automatic	0.0004
47	4	behavior	0.0142
63	2	bibliography	0.0026
11	1	command and control	0.0038
75	2	communication	0.0022
187	11	computer	0.0270
29	2	concept	0.0057
12	1	context	0.0035
11	2	control	0.0152
16	4	cybernetics	0.0417
33	2	decision	0.0051
16	1	engineering	0.0012
19	1	flow	0.0022
8	1	game	0.0052
4	1	geometr(y,ical,ic)	0.0104
87	2	grammar	0.0019
13	2	heuristic	0.0128
183	1	information	0.0002
69	3	language, artificial	0.0054
137	5	language, natural	0.0076
20	5	learning	0.0521
131	1	linguistics	0.0003
10	1	list processing	0.0042
78	3	logic	0.0048
73	3	machine	0.0051
35	1	man	0.0012
46	1	management	0.0009
4	1	Markov	0.0104
55	2	mathematic(s,al)	0.0030
36	1	memory	0.0012
77	4	models	0.0087
3	1	neuron	0.0139
5	2	neuropathology	0.0333
16	2	pattern	0.0104
21	6	problem-solving	0.0714
84	4	program(med,ming)	0.0079
15	1	Project MAC	0.0028
21	1	psychology	0.0020
37	2	questions	0.0045
22	2	recognition	0.0076
5	1	recursive	0.0083
151	2	retrieval	0.0011
73	2	review	0.0023
17	2	self-organiz(ation,ing)	0.0098
48	3	semantic(s)	0.0078
25	1	simulation	0.0017
47	1	statistic(s,al)	0.0009
77	1	syntax	0.0005
248	2	systems	0.0007
7	1	teaching	0.0060
105	4	theory	0.0063
6	2	thinking	0.0278
31	1	transformations	0.0013

Figure 1. Term Profile for ARTIFICIAL INTELLIGENCE

TOT. OCC.	CO-OCC.	TERM	ASSOC. COEFF.
46	1	behavior	0.0027
24	2	biology	0.0208
25	3	center	0.0450
134	1	chemistry	0.0009
73	1	communication	0.0017
46	3	costs	0.0245
71	1	design	0.0018
28	8	dissemination	0.2857
32	1	document	0.0039
34	3	documentation	0.0331
19	1	education (training)	0.0066
34	1	efficiency	0.0037
109	2	evaluation	0.0046
19	1	flow	0.0066
7	2	foreign	0.0714
24	1	format	0.0052
6	1	government	0.0208
73	1	index	0.0017
172	8	information	0.0465
13	1	input	0.0096
75	2	librar(ies,y)	0.0067
11	1	literar(y,ature)	0.0114
45	2	management	0.0111
29	3	medi(cal,cine)	0.0043
23	1	micro(form,film,card,image,fiche)	0.0054
4	1	operations	0.0312
49	3	organization	0.0230
25	1	patents	0.0050
31	3	periodicals	0.0363
21	2	psychology	0.0238
19	2	publication	0.0263
76	1	punched cards	0.0016
15	1	questionnaire	0.0083
4	1	report	0.0312
25	1	requirements	0.0050
150	2	retrieval	0.0033
73	2	review	0.0068
53	3	scien(ce,tific)	0.0212
2	1	Science Information Exchange	0.0625
29	1	scientists	0.0043
11	1	standards	0.0114
44	1	storage	0.0028
241	2	systems	0.0021
7	1	teaching	0.0179
49	1	use	0.0026

Figure 2. Term Profile for DISSEMINATION + INFORMATION

The associativity formula corresponds to the product of the conditional probabilities that, given a document which one term indexes, the other term is also used to index it. The formula was specifically suggested for use in the system by Stiles, and should not be confused with his formula based on the chi-square distribution, which he uses to calculate an association factor between pairs of index terms.

In the retrieval technique originally used in the experiment, the association coefficients served as weights for the profile terms in a second search of the document collection. The list of profile terms was compared with the index terms of the documents in the collection and the weights of the terms that matched were summed, resulting in a relevance number for each document. This document relevance number was used to present the documents in the order of their probable relevance to the request. Many of the concepts providing the theoretical foundation for this technique were expressed by Maron and Kuhns [4]. The notion of summing term associativity coefficients to produce document relevance numbers was shown to be effective on an existing collection of documents by Stiles [5], but a different method for calculating the weights of the associated term was employed.

To automatically expand the original output, the second pass was performed on the entire document collection. This was intended to enable the retrieval system to locate documents relevant to the original request even though those documents had not been indexed by the terms in the request. To automatically narrow the original output, only the documents retrieved by the initial request were considered. This was based on the assumptions that enough potentially relevant documents had already been retrieved and that the ranking method given above would present the documents most relevant to the request at the head of the list.

The experiment attempted to answer three basic questions:

1. Can the automatic techniques effectively expand the original search output?

2. Can the automatic technique effectively narrow the original search output?

3. Assuming an affirmative answer to the first two questions, can a threshold level be applied to the associativity coefficients which reduces the size of the term profile without influencing the results of the automatic techniques?

The basis for this experiment was the document collection of the Center for the Information Sciences at Lehigh University. This collection contains approximately 2,500 documents, which are manually indexed using a thesaurus composed of 450 index terms. A total of ten searches were performed on the document collection and provide the data on which the results are based.

In designing the experiment, the intent was to select a large enough sample of initial Boolean search statements to lend an acceptable degree of validity to the results. This was dictated, to a certain extent, by the practical restrictions imposed by economics and time, and by the very real consideration of the absence of a suitable definition for "a large enough sample." Despite the inevitable small sample size, it was felt that a certain amount of insight could be achieved through the selection of a realistic and representative group of search statements. The other limitation affecting the results of the experiment was the use of an intuitive basis for evaluation. This was born out of necessity, however, as no suitable alternative was apparent.

C. Experimental Results

Table 2 in the Appendix gives the results obtained from the initial search statement, "DISSEMINATION + INFORMATION," which retrieves every document in the collection indexed by the terms "dissemination" and "information." This request produces eight documents, which are denoted in Table 2 by an asterisk (*), and the term profile presented earlier in Figure 2. The remaining documents in Table 2 are produced by applying the associativity coefficients from the term profile to the automatic expansion procedure. The resulting highest document relevance numbers are shown in Table 2, along with the ranks and titles of the documents. It should be noted that no significance is attached to the document relevance number except as a basis for ranking the documents.

In looking at the documents produced by the expansion technique, there appear to be several documents which can be intuitively judged as highly related to the request, and therefore, of potential interest to the user who wants to examine more than the original eight. However, the search terms chosen were not the precise terms originally used to index these documents and, as a result, they were not retrieved. This is a very real problem in a coordinate indexing system using several hundred terms to index documents on various aspects of a particular subject. The indexer tries to use language he hopes will be used by future requesters and, to a certain extent, the user tries to use the language he thinks the indexer used, but in many cases they don't settle on the same set of terms.

Although the results illustrated by Figure 3 are encouraging, it must be remembered that the expansion was based on the indexing of eight documents. Since the validity of the information contained in the term profile increases with the amount of data available to generate it, the ability to automatically expand a search retrieving only one or two documents is questioned. In an operational retrieval system, expansion is probably most desirable when the original output is small, so to be of any practical value, the expansion technique must be effective when very few documents are initially retrieved.

TOT. OCC.	CO-OCC.	TERM	ASSOC. COEFF.
24	1	artificial intelligence	0.0417
13	1	automata	0.0769
182	1	computer	0.0055
11	1	control	0.0909
19	1	learning	0.0526
54	1	mathematic(s,al)	0.0185
77	1	models	0.0130
15	1	pattern	0.0667
22	1	recognition	0.0455
17	1	self-organiz(ation,ing)	0.0588
237	1	systems	0.0042
103	1	theory	0.0097

Figure 3. Term Profile for LEARNING + SYSTEMS

Table 3 in the Appendix represents the expansion of "LEARNING + SYSTEMS" based on the term profile shown in Figure 3. This request produced only one document, but the expansion appears to have produced some potentially relevant documents. It should be noted, however, that the document originally retrieved was indexed by twelve terms, slightly more than twice the average in the CIS document collection. This caused the generation of a larger term profile than would normally be the case when the original output is only one document. A smaller term profile could have resulted in less satisfactory expansion results.

Another problem caused by small initial output becomes apparent upon closer scrutiny of Figure 3. When a term profile is generated from a small number of documents, the formula for calculating associativity coefficients appears to place the index term that is quite frequently used in the total collection at an unfair disadvantage. For example, the term "systems" co-occurred in the maximum number of original documents, but received the lowest associativity coefficient because its total occurrence in the document collection was the highest of the twelve associated terms. However, it doesn't appear to be detrimental to the results of the expansion in this case.

In attempting to narrow the search automatically, the assumption is made that all the documents originally retrieved are potentially relevant to the user's request. It remains for the automatic technique to present these documents in their order of relevance to the request so the user can halt the retrieval process when he has enough documents to work with. The problem lies in determining the order of relevance of the documents to the request.

The technique was used to determine the order of relevance of the twenty-four documents retrieved by the search "ARTIFICIAL INTELLIGENCE." The term profile generated for this search is given in Figure 1 and the results of the document ordering are shown in Table 1 in the Appendix. It is obvious from this example that when a summation of associativity coefficients is all that is involved in producing document relevance numbers, the number of terms used to index a document becomes an extremely influential parameter. Since the term profile is composed of the terms indexing the documents originally retrieved, an increase in the number of index terms for one of these documents can only result in

an increase in the relevance number for that document. This relationship becomes quite significant when the document collection has been manually indexed, due to the well-known inconsistencies introduced into a system through the use of human indexers. Such is the case in the CIS system, where not one, but several indexers have been employed. Based on this shortcoming, it is evident that a modification must be made to the original automatic technique before it can effectively narrow a search.

It should be pointed out that the problem explained above has no bearing on the automatic expansion process examined earlier, as the concern there is with the documents not retrieved by the original search statement.

In spite of the unsatisfactory results involving the rankings produced by the document relevance formula, it was decided to continue to experiment with the use of a cut-off value for associativity coefficients. This called for using only profile terms with associativity coefficients greater than some predetermined threshold to calculate the document relevance number. The threshold chosen for the experiment was 0.0125. There was nothing magic about that particular figure. It was chosen because it had been used previously in the CIS on-line associative search program and generally produced a term profile of from ten to fifteen terms, which was considered a good number to work with for this experiment.

The use of the cut-off does not drastically affect the document rankings in any of the searches tested. This can be seen by the comparisons given of "DISSEMINATION + INFORMATION" and "ARTIFICIAL INTELLIGENCE" in Figure 4. The conclusions are that: 1) the highest document relevance numbers are primarily the result of a few terms with high associativity coefficients rather than several terms with low associativity coefficients, and 2) the weights of the terms deleted by the cut-off are not significant enough to cause any appreciable fluctuation in the document relevance numbers.

DISSEMINATION + INFORMATION				ARTIFICIAL INTELLIGENCE			
ORIG. RANK	ORIG. DOC.REL.NO.	C.-O. RANK	CUT-OFF DOC.REL.NO.	ORIG. RANK	ORIG. DOC.REL.NO.	C.-O. RANK	CUT-OFF DOC.REL.NO.
1	0.6320	1	0.5863	1	1.1858	1	1.1541
2	0.6106	2	0.5800	2	1.1527	8	1.0943
3	0.5413	3	0.4790	3	1.1467	2	1.1363
4	0.4669	5	0.4314	4	1.1419	4	1.1123
5	0.4563	4	0.4498	5	1.1339	"	"
6	0.4232	7	0.4172	6	1.1264	6	1.1107
7	0.4192	6	0.4192	7	1.1254	3	1.1254
8	0.3772	8	0.3772	8	1.1086	9	1.0933
9	0.3409	9	0.3322	9	1.1021	10	1.0799
10	0.3350	10	0.3307	10	1.0984	7	1.0984
11	0.3255	11	0.3220	11	1.0916	11	1.0714
12	0.3155	12	0.3095	12	1.0596	16	1.0270
13	0.3006	13	0.2857	13	1.0568	12	1.0559
14	0.2968	"	"	14	1.0521	13	1.0417
15	0.2965	"	"	15	1.0514	14	1.0409
16	0.2939	"	"	16	1.0490	18	1.0208
17	0.2911	"	"	17	1.0482	15	1.0333
18	0.2903	"	"	18	1.0327	16	1.0270
19	0.2892	"	"	19	1.0318	21	1.0000
20	0.2890	"	"	20	1.0275	19	1.0152
21	0.2878	"	"	21	1.0251	20	1.0139
22	0.2866	"	"	22	1.0194	21	1.0000
23	0.2857	"	"	23	1.0026	"	"
"	"	"	"	24	1.0000	"	"
"	"	"	"				
"	"	"	"				
"	"	"	"				
"	"	"	"				
29	0.1813	29	0.1702				
30	0.1503	32	0.1278				

Figure 4

This implies that the automatic technique being tested will still be effective in expanding a search when the cut-off is used. However, it remains to find a suitable modification to the document relevance formula before any degree of success is achieved in automatically narrowing a search. In both cases, the use of a cut-off will result in a considerable savings in computer time, owing to the sizable reduction in the number of index terms used.

It had been pointed out earlier that the modification to the document relevance formula was necessary to neutralize the inconsistencies introduced by the indexing process. Of primary concern was the depth-of-indexing variation among indexers. With only a summation of term associativity coefficients determining the degree of relevance of the retrieved documents to the original request, it was feared that some documents might achieve unwarranted high rankings because one indexer typically assigned more index terms than the others.

After much experimentation, it was determined to use the following modification for calculating the new relevance number for a document:

$$\text{DOCUMENT RELEVANCE NUMBER} = \frac{S \times N}{T}, \text{ where}$$

S = document relevance number as calculated previously, but using only terms with associativity coefficients ≥ 0.0125 in the summation.

N = number of terms with associativity coefficients ≥ 0.0125 indexing the document.

T = total number of terms indexing the document.

The affect of this formula can be seen in Figure 5. This shows the new ordering of the documents retrieved by "ARTIFICIAL INTELLIGENCE," and compares it to the ranking produced for the same search by the original and cut-off methods presented earlier. The total number of index terms and the index terms with associativity coefficients ≥ 0.0125 are also given for each document. In the latter column, "a.i." is used for the index term "artificial intelligence." The titles of the documents were given in TABLE 1 and the term profile for the search "ARTIFICIAL INTELLIGENCE" was shown in Figure 1.

While the cut-off method produced no major deviation from the original results, it is apparent that the new document relevance number formula has. The documents heading the list are generally those indexed by the greatest percentage of "significant" index terms (those which survived the cut-off). At the same time, the relationships among index terms, provided by the associativity coefficients, remain an important parameter and still influence the final document ranking. With the new formula, a document is not penalized for having a greater than average number of index terms if those terms are "significant."

ARTIFICIAL INTELLIGENCE

ORIG. RANK	C.-O. RANK	NEW RANK	NEW DOC.REL.NO.	NO. TERMS	TERMS ≥ 0.0125
1	1	9	0.5775	10	a.i., learning, neuropathology, computer, cybernetics
2	8	19	0.3648	12	a.i., control, computer, learning
3	2	4	0.9092	5	a.i., heuristic, learning, problem-solving
4	4	12	0.4944	9	a.i., problem-solving, answers, computer
5	4	12	0.4944	9	a.i., problem-solving, answers, computer
6	6	10	0.5554	10	a.i., behavior, computer, cybernetics, thinking
7	3	1	1.1254	5	a.i., computer, behavior, problem-solving, heuristic
8	9	8	0.6248	7	a.i., behavior, learning, computer
9	10	14	0.4629	7	a.i., thinking, learning
10	7	2	1.0984	3	a.i., problem-solving, computer
11	11	15	0.4286	5	a.i., problem-solving
12	16	23	0.2054	10	a.i., computer
13	12	5	0.7920	4	a.i., cybernetics, behavior
14	13	6	0.6944	3	a.i., cybernetics
15	14	20	0.3093	8	a.i., computer, ALGOL
16	18	21	0.2552	8	a.i., adaptive
17	15	16	0.4134	5	a.i., neuropathology
18	16	7	0.6846	3	a.i., computer
19	21	24	0.1429	7	a.i.
20	19	17	0.4060	5	a.i., control
21	20	18	0.4056	5	a.i., neuron
22	21	22	0.2500	4	a.i.
23	21	11	0.5000	2	a.i.
24	21	3	1.0000	1	a.i.

Figure 5

The use of this formula provides a major improvement in the rankings of the documents and therefore in the ability to automatically narrow the search. It should be noted that, for purposes of narrowing the search, this formula could not be used without the existence of a cut-off, since every index term in the documents originally retrieved appears in the term profile. So the cut-off is not merely important in saving computer time, it is mandatory if the automatic narrowing technique is to function properly.

Since it would be more convenient to use the same document relevance formula in both the automatic expansion and narrowing techniques, the new formula was tested with the automatic expansion technique, despite the fact that the technique had already proven to be effective.

Figure 6 illustrates the type of results produced by the use of the new document relevance number formula with the automatic expansion technique. The headings for this figure are identical to those used in Figure 5. The documents originally retrieved are denoted by an asterisk (*) before the index terms. The original expansion of the search "DISSEMINATION + INFORMATION" was given in TABLE 2 and the term profile was shown in Figure 2.

DISSEMINATION + INFORMATION

ORIG. RANK	C.-O. RANK	NEW RANK	NEW DOC.REL.NO.	NO. TERMS	TERMS ≥ 0.0125
1	1	6	0.2932	18	*dissemination, information, foreign, biology, periodicals, center, publication, documentation, scien-
2	2	2	0.3988	16	*dissemination, information, teaching, biology, librar-, costs, organizat-, documentation, patents, foreign, government
3	3	8	0.2579	13	*dissemination, information, scien-, organization, center, costs, documentation
4	5	16	0.1327	13	*dissemination, information operations, biology
5	4	4	0.3374	8	*dissemination, information, periodicals, psychology, publication, report
6	7	5	0.2980	7	dissemination, foreign, periodicals, psychology
7	6	1	0.4192	4	*dissemination, information, science info exchange, costs
8	8	3	0.3772	4	*dissemination, information, psychology, scien-
9	9	15	0.1329	5	*dissemination, information
10	10	9	0.2480	4	dissemination, scien-, psychology
11	11	18	0.1288	5	dissemination, periodicals
12	12	10	0.1857	5	dissemination, psychology, science info exchange
13	13	35	0.0319	9	dissemination
14	13	19	0.0952	3	dissemination
15	13	32	0.0408	7	dissemination
16	13	19	0.0952	3	dissemination
17	13	19	0.0952	3	dissemination
18	13	25	0.0714	4	dissemination
19	13	25	0.0714	4	dissemination
20	13	19	0.0952	3	dissemination
21	13	19	0.0952	3	dissemination
22	13	11	0.1429	2	dissemination
23	13	7	0.2857	1	dissemination
"	13	19	0.0952	3	dissemination
"	13	11	0.1429	2	dissemination
"	13	25	0.0714	4	dissemination
"	13	11	0.1429	2	dissemination
"	13	25	0.0714	4	dissemination
29	29	14	0.1418	6	information, organization, center, costs, operations
30	32	34	0.0383	10	information, periodicals, center

Figure 6

The change in rankings shown in Figure 6 is not too pronounced. The documents retrieved by the original expansion were also retrieved using the new formula, so in that sense the automatic expansion technique is still effective, because it still provides the user with potentially relevant documents he missed with his original search.

Since the problems encountered with the automatic narrowing technique have been solved by the change in the document relevance number formula, and the automatic expansion technique, for all practical purposes, performs as well with that formula as it did with those previously tested, it was determined that a satisfactory method had been found to both automatically expand and automatically narrow the output from the initial search statement.

In summarizing, the basic steps in the proposed automatic retrieval method for the CIS system are:

1. Generate a term profile from the user's initial Boolean search statement using the existing formula to calculate the term associativity coefficients.

2. Using only terms with associativity coefficients ≥ 0.0125, compare the list of profile terms with the index terms of each document and add the associativity coefficients of the terms that match. The sum, S, of the weights is used to calculate the document relevance number. To expand the original search, perform this step for every document in the collection. To narrow the original search, use only the documents originally retrieved.

3. For each document, multiply S from step 2 by the number of terms with associativity coefficients ≥ 0.0125 indexing that document. Then divide the product by the total number of terms indexing the document. The result is the document relevance number.

4. Present the documents to the user in the order of their probable relevance to his request.

D. Conclusion

The CIS document retrieval system must be prepared to serve many different types of users, each of whom may have different needs. Because of this, it is unreasonable to expect a one-pass search process to satisfy all user classes. This problem has been alleviated somewhat

in the system by the introduction of a semi-automatic multiple-pass search strategy. This procedure leaves the search reformulation in the user's hands and allows for different adjustments from user to user, depending on individual needs. However, there will still be times when the user is unable to optimize his search results by reformulating his own request. When this occurs, his needs may best be served by a fully automatic retrieval strategy in which his only function is to criticize the initial search as being too narrow or too broad. The results of this experiment indicate that the proposed automatic retrieval technique could be successful in that exact situation.

Another problem has been created in the CIS system by the addition of new index terms to the thesaurus. When a new term is added, it is not feasible to re-index the entire document collection based on that term. The addition of a new term could result in documents not being indexed by that term which, in fact, should be. Consequently, the possibility exists of potentially relevant documents not being retrieved when the new term is used in a search statement. It is believed the automatic expansion technique described in this paper could retrieve many of these documents, although not enough experimentation was performed to present any valid evidence to that effect.

The proposed technique lends itself quite easily to an on-line environment. Although the search procedure itself is completely automatic, the user maintains control with the capability of halting the operation any time he has enough documents to satisfy his needs. A conversational mode could be developed to offer the user a choice among a non-associative search, a semi-automatic associative search, and a completely automatic associative search. The first decision could be between non-associative and associative search. If the latter option was taken, a second decision would have to be made between the two types of associative searches.

In an off-line, batched processing environment, the proposed technique could still operate if the user knew prior to the run how many documents he wanted to retrieve. After the initial search had been completed, a comparison would be made between the number of documents retrieved and the number of documents the user wanted to retrieve. Based on this comparison, the program would either automatically expand

the original search, automatically narrow the original search, or stop searching altogether. The final operation would be the presentation of the exact number of documents the user specified in the order of their relevance to his initial request. Since no manual intervention is required, the user would not incur a delay in the processing of his search request.

In either of the aforementioned environments, the biggest problems in implementing the proposed technique are the limitations imposed by sorting the documents in relevance number sequence. A time-consuming tape-sort operation is considered prohibitive, especially in an on-line environment. It is not unrealistic to consider a core-sort for the automatic narrowing procedure, since there generally exists a relatively small, fixed number of document records being treated. However, an algorithm must be derived to reduce the number of document records being sorted before the automatic expansion technique can be handled in core. If this problem can be alleviated, it appears likely the proposed technique could perform within acceptable time limits in the CIS system.

Although the results presented in this paper are encouraging, in the last analysis, the best way to evaluate the effectiveness of any technique such as the one described is to observe its use in an operational environment.

References

[1] Arthur D. Little, Inc., Centralization and Documentation, Second Edition, C-64469, Cambridge, Massachusetts, June, 1964.

[2] Curtice, Robert M., and Rosenberg, Victor, Optimizing Retrieval Results with Man-Machine Interaction, Center for the Information Sciences, Lehigh University, Bethlehem, Pennsylvania, February, 1965.

[3] Curtice, Robert M., "Experiments in Associative Retrieval," Progress in Information Science and Technology, October, 1966.

[4] Maron, M. E., and Kuhns, J. L., "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of the Association for Computing Machinery, Volume 7, Number 3, July, 1960.

[5] Stiles, H. Edmund, "The Association Factor in Information Retrieval," Journal of the Association for Computing Machinery, Volume 8, Number 2, April, 1961.

APPENDIX: Ranked Document Listings

TABLE 1: ARTIFICIAL INTELLIGENCE
Ranked Search and Expansion

RANK	DOC.REL.NO.	DOCUMENT
1	1.1858	*Kochen, M., Mackay, D., Maron, M. "Computers and Comprehension." Santa Monica, RAND Corp. RM-4065-BR, AD 437 589, Apr 1964
2	1.1527	*Tou, J. and R. Wilcox Eds. Computer and Information Sciences. Spartan Books Inc, Washington, 1964
3	1.1467	*Gelernter, H. "Intelligent Behavior in Problem-Solving Machines." IBM J of Res and Develop 2, 336-345, Oct 1958
4	1.1419	*Raphael, B. "A Computer Program Which Understands." Fall Joint Comput Conf, Proc., 1964, 577-589
5	1.1339	*Bobrow, D. "A Question Answering System for High School Algebra Word Problems." Fall Joint Comput Conf, Proc., 1964, 591-614
6	1.1264	*Maron, M. "On Cybernetics, Information Processing & Thinking." Santa Monica, RAND Corp. P-2879, AD 435 484, Mar 1964
7	1.1254	*Hormann, A. "Computers in Behavioral Science." Behavioral Science 10, 88-107, Jan 1965
8	1.1086	*Strom, R. "Methodology for Research in Concept Learning." IBM Res Center, Yorktown Heights, NY, AFCRL-64-87, AF 19(638)-2752, Apr 1964
9	1.1021	*Kochen, M. "A Model for the Process of Learning to Comprehend." IBM Res Center, Yorktown Heights, NY, AFCRL-64-87, AF 19(638)-2752, Apr 1964
10	1.0984	*Hormann, A. "Three Branches of Artificial Intelligence Research." Syst Develop Corp, SP-1858/000/01, Nov 1964
11	1.0916	*Mesarovic, M. "Toward a Formal Theory of Problem Solving." In Symp on Comput Augmentation of Human Reasoning, Washington, 1964, Proc. Spartan, 1965, 37-64
12	1.0596	*Raphael, B. "SIR, A Computer Program for Semantic Info Retrieval." MIT, MAC-TR-2, June 1964
13	1.0568	*Beer, S. "The Biophysical Theory of Cybernetics." In his Cybernetics and Management, John Wiley and Sons Inc, 1959, 105-142

Preceding Page Blank

Table 1-2

RANK	DOC.REL.NO.	DOCUMENT
14	1.0521	*Ashby, W. An Introduction to Cybernetics. London, Chapman and Hall Ltd, 1956
15	1.0514	*Feldman, J. "Aspects of Associative Processing." Lincoln Lab, MIT, Tech Note 1965-13, AD 614 634, Apr 1965
16	1.0490	*Rome Air Dev. Cent. "The Study of Mathematical Models for Self Organizing Systems." Griffiss Air Force Base, NY, RADC-TDR-64-328, Final Rept, Jan 1965
17	1.0482	*Watt, W. "PLACEBO IV, Rules, Concordance, Sample Computer Generation." NBS, Tech Note 255, Mar 1965
18	1.0327	*Engelbart, D. "Augmenting Human Intelligence, A Conceptual Framework." Stanford Res Inst Proj 3578, AF 49(638)-1024, AFOSR-3223, Oct 1962
19	1.0318	*Murray, A. "Information Processing Relevant to Military Command Bibliography." L. G. Hanscom Field, Bedford, Mass, 2 Vols, AD 418 152, AD 418 176, Feb 1963
20	1.0275	*Pask, G. "Teaching as a Control Engineering Process." Contr and Automat Progr 9, 6-11, Jan 1965
21	1.0251	*Good, I. "Speculations Concerning the First Ultra-intelligent Machine." Advances in Computers, Academic Press, 1965, 31-88
22	1.0194	*Hughes Aircraft Co. "Creative Computation." Griffiss Air Force Base, NY, RADC-TR-65-123, June 1965
23	1.0026	*Minsky, M. "A Selected Descriptor Indexed Bibliography to the Literature on Artificial Intelligence." In Feigenbaum, E. and J. Feldman Eds, Computers and Thought, McGraw Hill, 1963, 453-523
24	1.0000	*Samuel, A. "Artificial Intelligence." Computer and Automation 12, 28-35, Mar 1963

APPENDIX: Ranked Document Listings

TABLE 2: DISSEMINATION + INFORMATION
Ranked Search and Expansion

RANK	DOC.REL.NO.	DOCUMENT
1	0.6320	*International Conf. Sc. Info. "Effectiveness of Monographs, Compendia, and Specialized Centers." Int Conf on Scien Inform, Proc. Washington, 1958, V.1, 541-659
2	0.6106	*International Conf. Sc. Info. "Responsibilities of Govt, Soc, Univ, + Ind for Improved Info Services." Int Conf on Scien Inform, Proc. Washington, 1958, V.2, 1415-1545
3	0.5413	*Jensen, R. "N.F.S.A.I.S. Proceedings for 1963 Annual Meeting." Nat Fed of Scien Abstr and Indexing Serv, Proc. Washington, March 1963
4	0.4669	*International Bus. Mach. "The IBM Data Systems Div Tech Info Center." IBM Data Syst Div, TR00.1103, Feb 1964
5	0.4563	*American Psych. Assn. "The Role of the Tech Rept in the Dissemination of Scientific Information." American Psychol Assn. APA-PSIEP No. 13, Apr 1965
6	0.4232	American Psych. Assn. "Proj on Info Exch, A Prelim Study of Info Exch Activities." American Psychol Assn. PSIEP Rept 10, June 1964
7	0.4192	*U. S. Senate. "Coordination of Information on Current Scientific Research and Development." 87th Congress Senate Comm on Govt Oper Rpt 268, May 1961
8	0.3772	*Garvey, W., Griffith, B. "Structure, Objectives, and Findings of a Study on Sci Info in Psych." Amer Doc 15, 258-267, Oct 1964
9	0.3409	*Ackoff, R., Halbert, M. "An Operations Research Study of the Diss of Scientific Information." Int Conf on Scien Inform, Proc. Washington, 1958, V.1, 97-130
10	0.3350	Griffith, B., Garvey, W. "Systems in Scientific Info + the Effects of Innovation + Change." Amer Doc Inst, Annual Meeting, 1964, Proc. 1, 191-200
11	0.3255	Case Institute. "Op Res Study of Dissemination and Use of Recorded Scientific Info." Case Inst of Tech G-8434, Dec 1960

Table 2-2

RANK	DOC.REL.NO.	DOCUMENT
12	0.3155	American Psych. Assn. Proj on Info Exch, "Theoret + Meth Considerations." Amer Psychol Assn. PSIEP Rpt 12, Jan 1965
13	0.3006	Lowry, W. and J. Albrecht. "A Proposed Info Handling System for a Large Research Organization." Int Conf on Scien Inform, Proc. Washington, 1958, V.2, 1181-1202
14	0.2968	Hindson, R. "The Dissemination of Published Information to the Executives of a Major Steel Group." ASLIB Proceedings 17, 8-22, Jan 1965
15	0.2965	Tritschler, R. "A Computer Integrated System for Centralized Info Dissem, Storage and Retrieval." Readings in Inform Retrieval, Scarecrow Press Inc, NY, 1964, 518-545
16	0.2939	Tauber, A., Meyers, W. "Photochromic Micro Images, A Key to Microdocument Storage and Dissemination." Amer Doc 13, 403-409, Oct 1962
17	0.2911	Microcard Corp. "Planning Guide for a Miniaturized Doc Distribution System." Microcard Corp., Dec 1962
18	0.2903	Henseley, C. "Selective Dissemination Pilot Study." IBM Res Center and Advanced Syst Develop Div, Jan 1961
19	0.2892	Martin, M., Ackoff, R. "Dissemination and Use of Information." Management Science 9, 322-336, Jan 1963
20	0.2890	Schultz, L. "RAPID, A System for Retrieval Through Automated Publication and Information Digest." Amer Doc Inst, Annual Meeting, 1964, Proc. 1, 79-87
21	0.2878	Koriagin, G. "Library Information Retrieval Program." Readings in Inform Retrieval, Scarecrow Press Inc, NY, 1964, 546-568
22	0.2866	Arverson, M. "Economic Aspects of Dissemination of Chemical Knowledge." J of Chem Doc, Nov 1961
23	0.2857	Davison, R. "An Announcement and Request for Initial Dissemination." Amer Doc. Inst, Annual Meeting, 1964, Proc. 1, 111-115
"	"	Kochen, M., Flood, M. "Some Bibliographic and Sociological Devices to Improve Maintenance of Current Awareness." IBM Res Center, Yorktown Heights, NY, AFCRL-64-87, AF 19(638)-2752, Apr 1964

Table 2-3

RANK	DOC.REL.NO.	DOCUMENT
"	"	Luhn, H. "A Business Intelligence System." IBM J 2, 314-319, Oct 1958
"	"	Luhn, H. "Automated Intelligence Systems - Some Basic Problems and Prerequisites for Their Solutions." Clarification, Unification, and Integration Stor and Retr, Proc, NY, Feb 1961, 3-20
"	"	Luhn, H. "Selective Dissemination of Information." Amer Doc 12, 131-138, Apr 1961
"	"	Resnick, A. "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents." Science 134, 1004-6, Oct 1961
29	0.1813	Overmyer, L. "Test Program for Evaluating Procedures for Exploitation of Literature." Western Reserve Univ, NSF-G-10338
30	0.1503	Sharp, H. Ed. "Need for Information." Readings in Inform Retrieval, Scarecrow Press Inc, NY, 1964, 17-86

APPENDIX: Ranked Document Listings

TABLE 3: LEARNING + SYSTEMS
Ranked Search and Expansion

RANK	DOC.REL.NO.	DOCUMENT
1	0.4840	*Tou, J. and R. Wilcox Eds. Computer and Information Sciences. Spartan Books Inc., Washington, 1964
2	0.1752	Spangler, M. "General Bibliography on Information Storage and Retrieval." General Electric, Tech Inform Series, R62CD2, Mar 1962
3	0.1712	Kochen, M. "A Model for the Process of Learning to Comprehend." IBM Res Center, Yorktown Heights, NY, AFCRL-64-87, AF 19(638)-2752, Apr 1964
4	0.1539	Murray, A. "Information Processing Relevant to Military Command Bibliography." L. G. Hanscom Field, Bedford, Mass, 2 Vols., AD 418 152, AD 418 176, Feb 1963
5	0.1534	Mitre Corp. "Self Organizing and Adaptive Information Systems." Mitre Corp., 1st Congress on Inform Syst Scien, Mitre SS-6, Nov 1962
6	0.1399	Schutzenberger, M. "On Probabilistic Push Down Storages in Self-Organizing Systems." M. Yovits, et al., Spartan, 1962, 205-213
7	0.1362	Rome Air Develop. Cent. "The Study of Mathematical Models for Self Organizing Systems." Griffiss Air Force Base, NY, RADC-TDR-64-328, Final Rept, Jan 1965
8	0.1326	Pask, G. "Teaching as a Control Engineering Process." Contr and Automat Progr 9, 6-11, Jan 1965
9	0.1223	Windknecht, T. "Concerning an Algebraic Theory of Systems." Case Inst Syst Res Center, AD 623 723, 1965
10	0.1177	Kirsch, R. "Computer Interpretation of English Text and Picture Patterns." IEEE Trans on Elec Computers, EC13, 363-376, Aug 1964
"	"	Stein, E. & Assoc. "Factors Influencing Design of Original-Document Scanners for Input to Computers." NBS Tech Note 245, Aug 1964
12	0.1136	Benington, H. "Military Information Recently and Presently." Mitre Corp., 1st Congress on Inform Syst Scien, Mitre SS-2, Nov 1962

Preceding Page Blank

Table 3-2

RANK	DOC.REL.NO.	DOCUMENT
"	"	Rome, B., Rome, S. "LEVIATHAN, An Experimental Study of Large Organizations with the Aid of a Computer." Mitre Corp., 1st Congress on Inform Syst Scien, Mitre SS-7, Nov 1962
14	0.1128	Kochen, M., Mackay, D., Maron, M. "Computers and Comprehension." Santa Monica, RAND Corp. RM-4065-BR, AD 437 589, Apr 1964
15	0.1122	Barus, C. "Scheme for Recognizing Patterns from Unspecified Classes." Swarthmore Coll, NSF G-5945, Dec 1961
"	"	Freeman, H. "Classification and Recognition for Geometric Patterns." NYU, Dept of Electrical Eng, Tech Rpt 400-33, July 1961
"	"	Rabinow, J. "Optical Character Recognition Today." Data Processing Mag. 8, 18-24, Jan 1966
"	"	Taube, M., Jones, R. "Distinction Between Character Recognition and Perceiving Machines." Amer Doc 12, 292-293, Oct 1961
"	"	Uhr, L. "Pattern Recognition." In Electronic Information Handling, A. Kent et al., Spartan Books, 1965, 51-72
20	0.1093	Windknecht, T. "Concerning an Algebraic Theory of Systems." Case Inst, Systems Res Center, AD 623 723, 1965
21	0.0998	Strom, R. "Methodology for Research in Concept Learning." IBM Res Center, Yorktown Heights, NY, AFCRL-64-87, AF 19(638)-2752, Apr 1964
22	0.0951	Congress on the Information System Sciences, Session 1. "Concepts of Information." Mitre Corp, 1st Congress on Inform Syst Scien, Mitre SS-1 Nov 1962
23	0.0943	Gelernter, H. "Intelligent Behavior in Problem-Solving Machines." IBM J of Res and Develop 2, 336-345, Oct 1958

A SYNTACTICALLY ORIENTED NATURAL LANGUAGE DOCUMENT RETRIEVAL SYSTEM WITH A BROWSABILITY FEATURE

by

Andrew J. Kasarda

Abstract

This paper is concerned with the design and construction of the retrieval component of a document retrieval system. A text processing scheme is defined for syntactic and semantic reduction of full text. A retrieval model is defined and constructed in such a way as to be compatible with the document characterization process described. The use of natural language communication is provided for the inquirer and the system's browsability capability is described.

This research was supported by the National Science Foundation under Grants No. GN-451 and No. GE-2569

Table of Contents

Page

PRELIMINARIES 2-1

Introduction 2-1

General Text Processing Scheme 2-1

STRUCTURE OF THE RETRIEVAL COMPONENT 2-4

A Formal Retrieval Model 2-4

The Induced Retrieval Model 2-5

IMPLEMENTATION OF THE INDUCED RETRIEVAL MODEL 2-11

The Retrieval Scheme 2-11

File Organization 2-14

Retrieval System Flow Diagrams 2-17

SUMMARY 2-21

APPENDIX I 2-24

APPENDIX II 2-28

APPENDIX III 2-32

List of Figures

Page

1. A Generalized Retrieval Component Scheme 2-4

2. The Partition and Structure of the Document Space D Induced by the Document Characterization Process at the Document Level 2-6

3. The Partition and Structure of the Document Space D Induced by the Document Characterization Process at an Intermediate Structure Level 2-7

4. The Graphic Structure of the Document Space D Induced by the Document Characterization Process 2-9

PRELIMINARIES

Introduction

The field of Information Storage and Retrieval is concerned with the collection, organization and retrieval of recorded information, and recently, the means by which these processes can be automated.

This paper will be concerned with the automation of the retrieval component of a document retrieval system and in particular, with the description of a natural language man-machine communication scheme which will provide a browsability feature for the user. The feasibility of automating such a system will also be discussed.

General Text Processing Scheme

The theory of a natural language query scheme for an automated document retrieval system requires that the system be compatible with the text processing scheme used to describe or characterize the document collection. Hillman [1] states that

> "A theory of document retrieval is a deductive system of the operations governing the retrieval of those documents whose representations contain characteristics (index terms) judged to be relevant to the terms of a query."

To satisfy this requirement, it will be assumed that the document collection has been characterized by the automatic syntactic text processing system developed by Hillman and Reed [2]. An abridged description of this text processing system will be given at this point to provide the reader with some degree of familiarity with the system.

Document Characterization. A major hypothesis of the theory of text processing is that the characteristics assigned to a document give some indication of what it is that the document is about. This aboutness is regarded as an *a priori* matter of logic, semantics and syntax.

In order to determine what a document is about, a scheme was devised to identify the topic-denoting expressions occupying referential position within the sentences of a document. In English, this

referential position is occupied by the noun phrase. Thus, the text processing system examines the sentences of each document contained in a document collection and identifies the noun phrases in each sentence of the text. This process uses a computational scheme of syntactic analysis of text developed by Hillman and Reed [3]. It is based on a context-sensitive computational grammar which makes use of a limited dictionary look-up. The dictionary consists of about three hundred functor words and suffixes. Appendix I gives a listing of the items in the dictionary. The analyzer assigns a syntactic category to each word of the input document text and identifies nominal, prepositional and infinitive phrases. The analyzer also segments the input sentences into micro-sentences, that is, into syntactically simple sentences. This initial step in text processing is called "micro-categorization." The next step in document characterization is a process termed "macro-categorization." This is a method by which the topic-denoting expressions and their predicates are identified. The process consists of two steps, the first of which consolidates the microcategories into larger units called "macrocategories." In the second step, the topic-denoting expressions (potential document characteristics) and their predicates are isolated. These topic denoting expressions are the keys to the documents since they reference the major topics about which the documents make assertions. Appendix II gives some examples of input and output of the microcategorization and macrocategorization processes in this text processing system.

The final step in the processing of the text deals with assigning a measure or weight to each document characteristic and the process of vocabulary control. After macrocategorization of the text has been completed, the document characteristics are merged and sorted into alphabetic order. Like characteristics are combined and counted. Next, a measure of term-document connectivity is assigned to each term-document pair. This is done using the notion of "lines of connection" described by Hillman [4] and Goodman [5]. If the predicates are thought of as relations and the document characteristics their arguments, then a characteristic term t will have n lines of connection to a predicate P if P is an n-place predicate. Similarly, a characteristic term t will have m lines of connection to a document D if m is the sum of all lines of connection between t and the predicates P of

document D in which t appears as an argument. The latter is clearly an assumption of linearity since a characteristic term t will have 8 lines of connection to document D if t appears as, for instance, an argument of a four-termed predicate P_1, a three-termed predicate P_2 and a one-termed predicate P_3 all of which occur in document D. A term-document matrix, called the affiliation matrix is set up with entries corresponding to the lines of connection between documents and characteristics. This matrix is then multiplied by its transpose and the resulting matrix is a term-term matrix called incidence matrix. Its entries establish connections between characteristics via some document and is a measure of first-level connectivity for an n-termed predicate. The incidence matrix is then partitioned into its components called transition matrices and each is normalized. These transition matrices represent distinct genera of terms and hence, are highly associated with each other. Finally, from each transition matrix, a unique probability vector is extracted and normalized. This vector consists of those characteristics occupying the most central positions in a genus. The result of this text processing scheme produces for each document a set of characteristics which identify the major topics referred to in the document. The weight of each characteristic in a document provides a measure of its association with the document, and a given characteristic will usually have different weights relative to different documents. And no less important, the genus structure provides a powerful tool to be used in retrieval of documents from the collection. Appendix III provides an illustration of this automated process.

The Document Corpus. In order for a document retrieval system such as the one being proposed here to be realistic, document collections containing 100,000 documents or more would be appropriate. The measure of connectivity between characteristics and documents would best reflect the behavioral regularities inherent in such large collections. However, it is not necessary, at least initially, to have such a large document collection. A collection of documents such as that contained in the Center for the Information Sciences (C.I.S.) collection would certainly suffice. This is a fairly homogeneous non-static collection of about 2,500 documents treating a wide variety of topics related to Information Science and Retrieval.

STRUCTURE OF THE RETRIEVAL COMPONENT

A Formal Retrieval Model

Once the document characterization scheme has been selected for the document retrieval system, care should be exercised in constructing an appropriate retrieval component for the system.

In very general terms, the retrieval component of a document retrieval system can be thought of as consisting of a document space D, a retrieval prescription space P, and a mapping or transformation T which transforms a prescription from P into a set of documents from D.

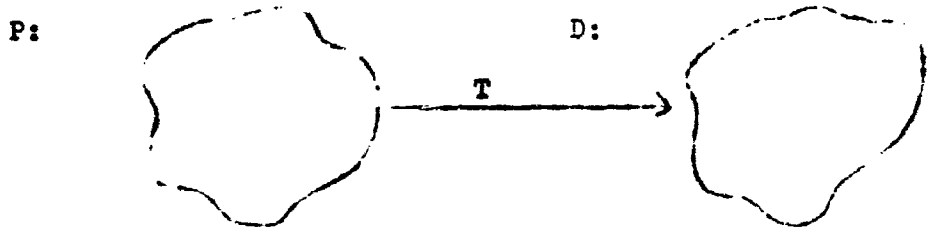


Figure 1. A Generalized Retrieval Component Scheme

The process of formally defining the retrieval component usually consists of imposing some kind of mathematical structure on D and P and then defining a transformation T in such a way that when the transformation T is applied to a retrieval prescription it will produce a relevant set of documents from D. However, the mathematical structures imposed on the document space D and the prescription space P are usually imposed on D and P without regard to the document characterization component of the document retrieval system.

For example, if a document collection C is the basis of a document retrieval system, the document space D can be thought of as consisting of all possible subsets of documents formed from the collection C. Note that if C contains n distinct documents, then the space D would consist

of 2^n distinct subsets of documents. Since the document space D is the power set of a finite aggregate of documents, an obvious structure for D immediately suggests itself. That is, the document space D is a finite Boolean algebra. The prescription space P is often given a more complicated structure that depends on the kind of index term to be used in P. If R is a repertory of simple descriptors, Mooers [6] points out that the descriptors can be thought of as two-element partially ordered systems where either the descriptor A is asserted as providing a clue to the document message, or no assertion, one way or the other, is made about A. The space P is simply the cardinal product of the two-element partially ordered systems in the repertory R. This space P of the retrieval prescriptions based on simple descriptors is a Boolean lattice. Each point x in P is a subset of descriptors from the repertory R. Given any such point x in the space P, there is a large family X of other points in P, each of which is "preceded" by the point x. Considering now the space D, there are many documents whose assigned subset of descriptors is one of the points belonging to the family X. If x^* is the largest set of documents which are indexed by a subset of descriptors in X, then the transformation T of the point x in P is the point x^* in D. The transformation T is the basis of selection in actual document retrieval systems based on descriptors.

It is quite clear that the mathematical structures imposed on the document space D and the prescription space P are compatible and also that the transformation T is well-defined. However, there is no reason to expect that the abstract mathematical structures imposed on D and P are, in fact, the real structures induced in D and P by the document characterization process. Therefore, the transformation T cannot be expected to be very efficient in the actual document retrieval process.

The Induced Retrieval Model

From the discussion above, it is clear that a formal approach to the construction of the retrieval component for a document retrieval system is not very useful. Since a document characterization process has been selected for the retrieval system, the obvious starting point in determining an appropriate retrieval structure lies with the document characterization process itself.

It seems quite reasonable to suspect that the structures induced in the document collection by the document characterization scheme should provide a clue to what features should be incorporated into the retrieval component to make it compatible with the other system components.

The document characterization scheme of Hillman [7] described earlier in this paper utilizes both syntactic and semantic analysis of full text, along with certain matrix operations and Markov processes to isolate sets of highly connected document characteristics. As a result, the document characterization process induces a partition of the document space D into mutually disjoint sets of connected documents. For instance, consider the result of the document characterization process applied to the document collection C given in the example in Appendix III.

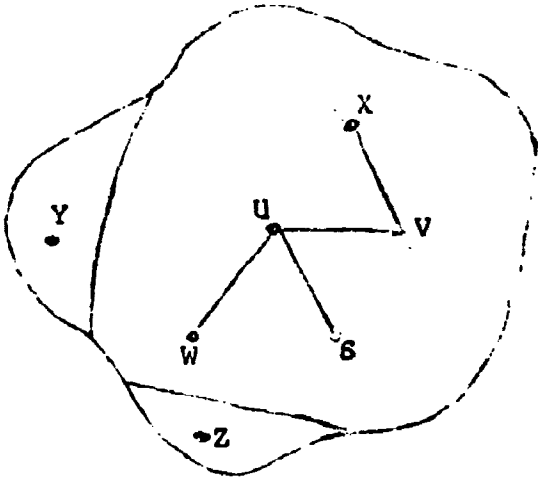


Figure 2. The Partition and Structure of the Document Space D Induced by the Document Characterization Process at the Document Level.

The structure of D at the document level clearly shows the relationships that do exist between the various documents in this trivial example, although nothing can be determined about the nature and the relative strength of the connections between the related documents. Thus, it appears that a more detailed picture of the structure of the

document space D is required if any insight is to be gained about the structure of the appropriate retrieval model. This can be done by simply adding the document characteristics to their respective documents and modifying the picture in Figure 2 in the obvious manner.

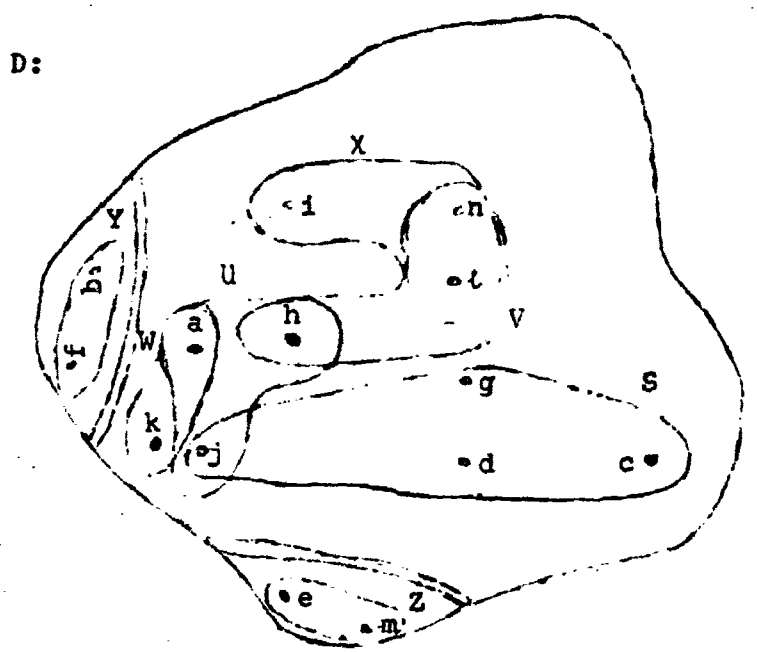


Figure 3. The Partition and Structure of the Document Space D Induced by the Document Characterization Process at an Intermediate Structure Level

At this intermediate level, the picture becomes a little clearer. It is now possible to define the distribution of the document characteristics between documents in the document space D. Note that for this particular example, the document U seems to tie the other documents in the genus together, and if it were removed, the document space D would be broken down into five distinct genera rather than the three genera defined in Figure 3. The document characteristics a, h, and j defining the document U are called articulation points of a first-level genus.

Another point, one that has important relation to the construction of the retrieval component, is that not every retrieval prescription in the prescription space P will have its image defined in D. There are several reasons for this situation. The most pronounced case occurs when a prescription contains document characteristics belonging to two or more genera. For instance, if the prescription were composed of the document characteristics b, a, and h, then no document or set of documents in D could satisfy the prescription. The result would be the empty set of documents. The other situations that would result in the empty document set are the obvious ones, foreign document characteristics, non-existent documents (within the collection), etc. In all of these situations, the retrieval system must be capable of guiding an inquirer by providing relevant clues via inquirer-retrieval system interaction in reformulating his original request. This will necessarily make the response transformation T quite complicated, in fact, it is very likely that no single transformation will suffice.

It is now quite apparent that the structure induced on the space D by the document characterization process is much finer than the Boolean algebra structure of P(C). The elements in the induced structure are, of course, also elements in P(C). But the elements of the induced structure in D are a very small "select" subset of the power set P(C). What the mathematical structure will be for the induced structure is difficult to say. Further investigation in this direction is being presently carried on by the author. However, it is not necessary to know what the mathematical structure is in order to construct the retrieval component. The induced structure provides sufficient information.

The document characterization process permits still another refinement of the document space structure. It is possible to interpret the incidence matrix (see Appendix III) as a graph G over the document characteristic repertory R. By the graph G(R) is meant that a given vertex set R of document characteristics forms a family of associations

$$A = (P_i, P_j)$$

where $P_i, P_j \in R$, indicating which vertices are connected characteristics. Each association A is called an edge of the graph, and the

multiplicity of an edge, denoted by

$$\mu(P_i, P_j) = \mu((P_j, P_i)$$

is simply the number of edges connecting the vertices P_i and P_j. Since a document characteristic is connected with itself, there is always an edge between any vertex and itself. Applying these notions to the example of Appendix III we get the following view of D.

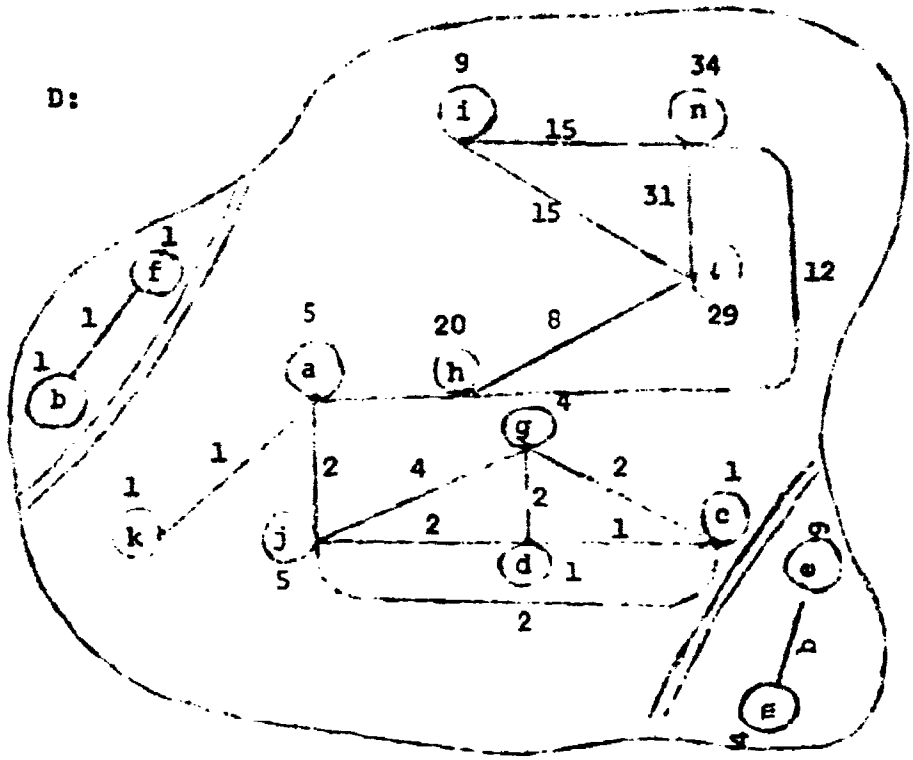


Figure 4. The Graphic Structure of the Document Space D Induced by the Document Characterization Process.

At this level of refinement, documents qua documents lose their identity and the structures discernable in the space D are only those structures defined on the document characteristics within the various genera. In essence, this is the concept level of the document space D and, in a sense, of the retrieval component itself. Intuitively, a prescription is simply a set of characteristics selected by an inquirer

as defining, in his mind, a concept (or concepts) about which he is interested. Actually the prescription defines what might be termed a "proto-concept" since the inquirer might not know precisely what it is that he wants. The selected characteristics then only provide the retrieval system with a clue to the actual concept being sought.

This seems to suggest that the document space D is really a subspace of the prescription space P. Since the inquirer's input language is to be his own natural language, the retrieval system should then be capable of inducing the same sort of structure in the prescription space P as the document characterization process induces in the document space D. Therefore, the retrieval component will require the same sort of syntactic and semantic analysis scheme on the natural language prescription to isolate the "proto-concept" characteristics in the prescription as was employed in the document characterization process.

The final step in the construction of the retrieval component is to describe the retrieval response transformations. They will be a family of transformations that will perform the following tasks:

1. Define the inquirer's prescription via direct inquirer-retrieval system interaction. This is the process of transforming the inquirer's proto-concept into a concept defined in the retrieval system;
2. Expand or narrow the retrieved document set;
3. Permit browsability within either documents or concept structures.

Transformations which will effectively perform the retrieval operations defined in (2) are respectively the double psuedo-complement and the double Brouwerian complement operators as described by Fairthorne [8]. They can be defined as follows:

Def.: The double-psuedo-complement, A**, of a set A is the smallest set that contains all documents indexed by A, but not only by A.

Def.: The double Brouwerian complement, 77 A, of a set A, is the largest set of documents containing nothing but A, but conceivably not all of the documents containing A.

The transformation that will permit browsability is one which, in effect, maps a characteristic into some genus at a given point and produces those sentences in a particular document (or a set of documents) in which the given characteristic occurs.

The process of defining the inquirer's prescription is a more difficult one and it requires a high degree of inquirer retrieval system interaction. The inquirer's response initially to a matching process in which a genus is located as being relevant to at least one of the terms of the prescription. The retrieval system's response is a set of connected terms from a genus that are closely associated with at least one of the terms in the prescription. The inquirer then responds with a reformulation of his original request and the process is repeated. This interaction then isolates or redefines the inquirer's prescription in terms compatible with the retrieval system without necessarily modifying his proto-concept.

Now that the structure of the retrieval component has been determined, all that remains is to describe its implementation.

IMPLEMENTATION OF THE INDUCED RETRIEVAL MODEL

In this section, a natural language retrieval system and its various related components will be described.

The Retrieval Scheme

It is now possible to describe a retrieval scheme that will permit natural language man-machine communication in a document retrieval system, as well as browsability in that system.

In a document retrieval system it must be agreed that generally such a system should provide documents as output to a given inquiry. However, the system will not be expected to provide facts as answers to a query. Therefore, this document retrieval system will not accept as valid any queries which are of an interrogative nature. That is, it will reject all queries in the form of a question. The reason for this restriction on the query structure is that questions usually require facts as answers, and since this is a document retrieval system and not a fact retrieval system, questions will be rejected as queries.

The browsability feature is made possible in this system as a result of the text processing scheme described earlier. The associations generated by this process permit association maps to be constructed for the various genera defined by the document collection, over which a query map is superimposed to isolate any relevant documents. As a further convenience to the inquirer, the browsability feature permits to scan the text of any document which he feels is interesting. The system displays various central topics sentences of those documents requested by the inquirer. He can narrow or expand this display simply by requesting that this be done.

A valid form of prescription is any sentence describing the topics which the inquirer is interested in. For example, a valid prescription could be the following:

> "I am interested in learning something about lattice structures and their relation to document retrieval systems. I am particularly interested in lattice structures and their relation to the retrieval component of a document retrieval system."

This prescription is, for all practical purposes, a demand made by the inquirer on the retrieval system to produce any documents which are about lattice structures and document retrieval systems. More precisely, about lattice structures and the retrieval component of a document retrieval system. The result of the syntactic analysis of the retrieval prescription would be just those noun phrases underlined in the discussion above. Since each sentence in the prescription would be treated as a distinct document, a connectivity structure for the prescription would be induced by the prescription (document) characterization process when applied to the prescription. The prescription characteristics along with their respective connectivity measures are given below.

prescription characteristic	connectivity measure
lattice structures	16/32
document retrieval systems	8/32
retrieval component of a document retrieval system	8/32

This induced structuring of the prescription permits a definite ordering or ranking of the output documents as implied by the inquirer in his prescription.

There are of course other methods of deriving "index terms" from the prescription. For example, Luhn's [9] KWIC indexing scheme could be applied to select the keywords in the prescription. However, there are two important disadvantages in using the KWIC indexing scheme. In the first place, the KWIC index terms probably would not be compatible with the system's document characteristics. Secondly, no implicit ranking scheme can be derived from the prescription for use in output ranking of retrieval documents. Thus the syntactic analysis approach [10], at least for this retrieval system seems most appropriate.

The output from the syntactic analysis of the query becomes the search request for first-level document search. First, an attempt is made to determine if the query characteristics are in the document characteristic table. The matching scheme used here should have the list processing capabilities inherent in either the LISP [11] or SNOBOL [12] compiler languages. The reason for such a capability is to permit the system to match single portions of the query characteristic with the document characteristics. This is necessary since a query characteristic would rarely have the same word composition as the document characteristic. The matching scheme isolates the various genera connected with the query characteristics. Once this is known, the system determines the most closely related terms in the respective genus and displays them to the inquirer. This affords the inquirer the capability of selecting actual document characteristics which are relevant to his request. The document characteristics which the inquirer feels are most relevant to his need are then re-entered into the system to be used in document selection. If the query characteristics did happen to match directly with various document characteristics, the system would proceed to select relevant documents.

At this point, the system determines the relevant documents first by scanning the Inverted File which lists those documents referenced by the particular query characteristic (actually this is a document characteristic). These documents are then selected from the Serial File and ordered in decreasing order of relevance to the query. This

can easily be done since in the Serial File, each referencing characteristic is listed with its associated connectivity value relative to that document. The document having the most relevant query characteristics and hence the highest connectivity value would be selected as most relevant to the query. Ordering of the total connectivity values will, in effect, order the retrieved documents. The retrieved documents are then displayed to the inquirer for approval. Here author-title information is displayed in place of the actual document. The inquirer can then decide on which documents interest him, if any at all. If too many documents were retrieved, the inquirer can narrow his search by re-selecting more characteristics from the relevant genus. If too few documents were selected, the inquirer can expand his search by deleting less important characteristics. In either case, the retrieval operations described above are repeated. Thus, the document retrieval system is quite flexible in its ability to accept queries, which are in a sense, simple association maps and superimpose them on the detailed association maps defined within each genus.

The retrieval system also provides the inquirer with the ability to browse through actual document text if he so desires. When the inquirer has found a document (or documents) which is of particular interest to him, he may ask the system to display for him those passages of text which are most central to the document. This is possible since the document characteristics are actual phrases appearing in the text and are central topics treated in the document. Therefore, the inquirer can simply ask the system to display those passages in a particular document by listing the document characteristics which are most appropriate to his needs. The system will display these passages from the document text file, which consists of such passages rather than full text.

File Organization

The system will require several kinds of file structures designed to satisfy its various kinds of data files. These will now be described in some detail.

Document File. The C.I.S. document collection will be used as the document data base. These documents will be in machine-readable form and will reside on magnetic tapes as an off-line component of the file.

Serial File. The serial file will be a disk-oriented file composed of variable length unblocked records having the following format:

Doc. No.	Doc. Characteristics and their respective Connectivity Values	Author-Title Information

The variable length unblocked record scheme was chosen because it provides the most economic use of the disk. A record key is generated to describe the record structure used in each record and is stored as part of that record. This structuring depends on the disk storage device being used by the system.

The document number will be a disk address that is derived from the record key assigned to the block structure. A mathematical transformation or algorithm generates a numerical address at which the record is stored. No index is required to determine the location of any record in the file. Generally, however, extensive analysis of the record key structure and range is necessary to implement such a randomly organized file. The ideal routine for record key conversion produces a unique storage address for every record in a file. In this case a unique storage address is possible since there are only 2,500 such records. A simple conversion method that is flexible and easy to implement is the divide-remainder technique, whereby a record key is divided by a number selected to produce a quotient and a remainder. The quotient is discarded, and the remainder becomes the disk record storage address. The divisor should be either a prime number or a number ending in 1, 3, 7, or 9. This usually results in relatively few duplicate remainders, and thus relatively few address synonyms. The logical choice for a divisor is a prime number slightly less than the number of storage units allocated for the given data file. Tentative choices for a divisor may be tested in the key transformation routine to determine which produces the fewest synonyms.

The second part of the document record (serial file) consists of

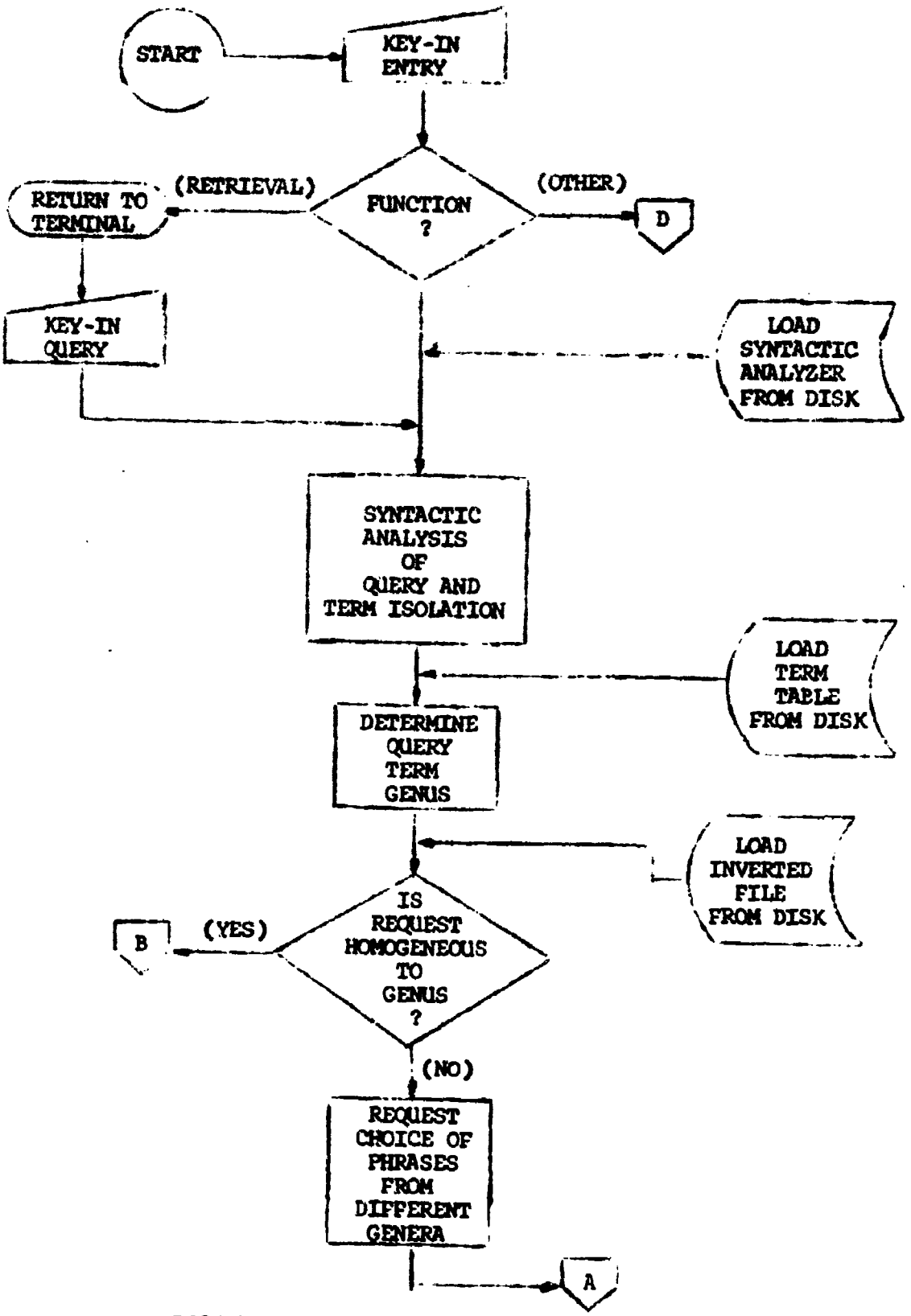
document characteristic numbers along with their respective lines of connectivity value for the given document. Each document will be allowed up to 10 document characteristics, each characteristic consisting of an average of 3 English words.

The third section of the document record is made up of author-title information to be used in retrieval returns in place of document numbers.

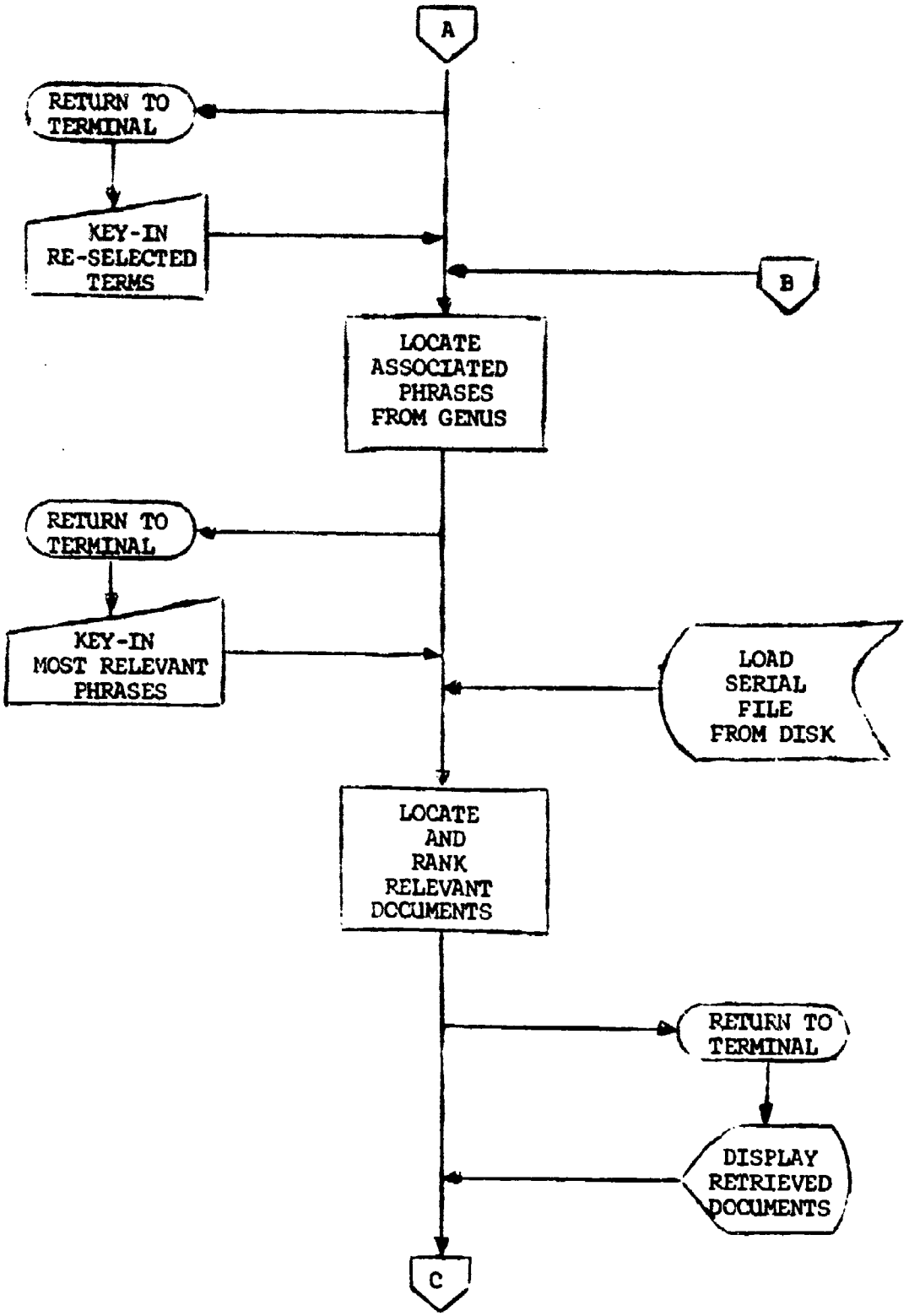
Inverted File. This file is a disk-oriented term-document file in which each record consists of a term number and the affiliated document numbers of those documents characterized by the given term. The characteristics are organized into genera which were defined during the initial text processing of the document collection. In this case, the file is in fixed-length blocked record format. A key provides the location of each genus in the file for fast access of these segments of the file. The inverted file will probably consist of around 5,000 document characteristics.

The Dictionary. The dictionary consists of the functors and suffixes that are used by the syntactic analyzer. It will also reside on the disk in a fixed-length blocked record format.

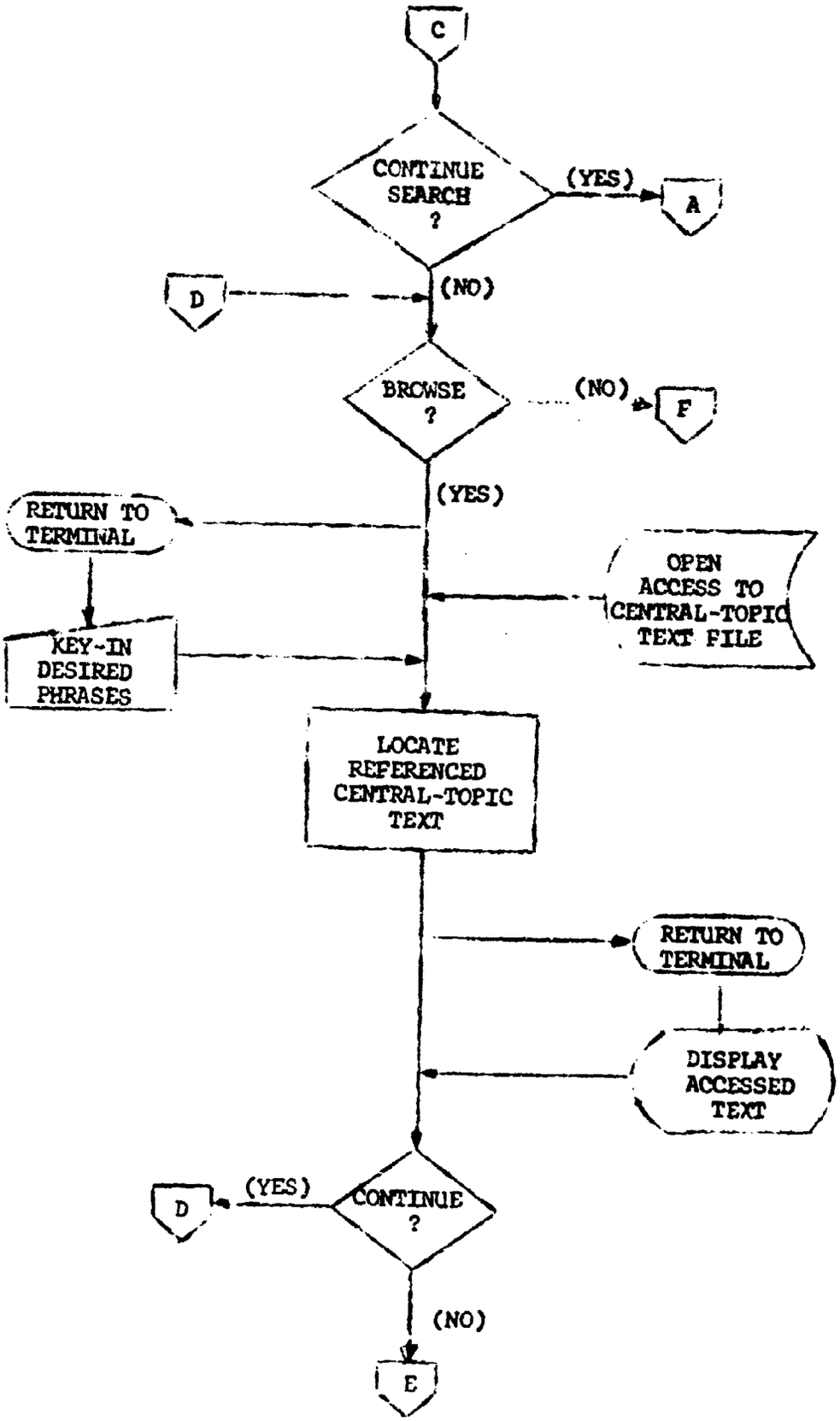
Systems Programs. All of the systems programs will reside on the disk and will be called by a program monitor that will reside in core memory. The programs are those required by the document retrieval system such as the syntactic analyzer, general retrieval programs, browsing programs, and system update programs. The following diagrams display the flow of the retrieval operations and the various options which are possible.



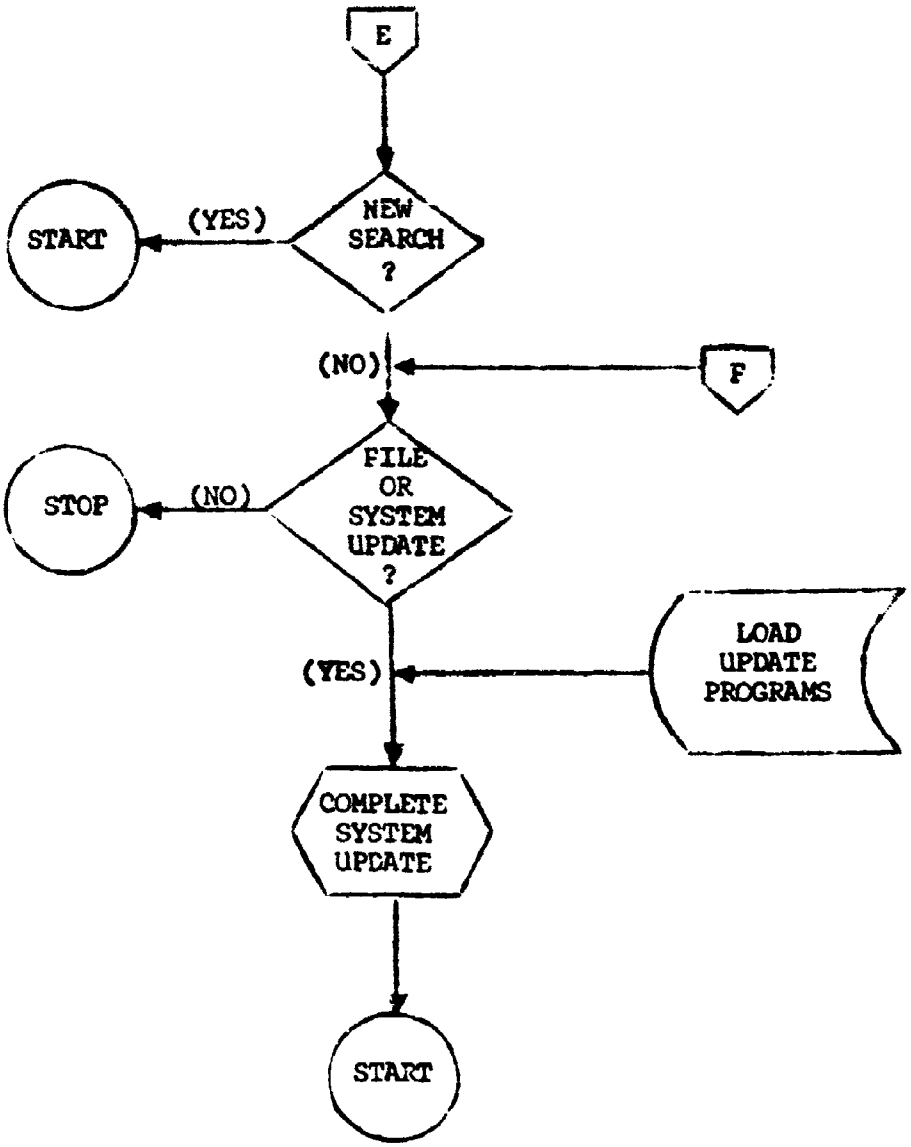
DOCUMENT RETRIEVAL OPERATIONS - (PART-A)



DOCUMENT RETRIEVAL OPERATIONS - (Part B)



DOCUMENT RETRIEVAL OPERATIONS - (Part-C)



DOCUMENT RETRIEVAL OPERATIONS - (PART-D)

SUMMARY

In the foregoing discussion, an attempt was made to construct a document retrieval system capable of natural language communication via man-machine interaction. The document base selected was the C.I.S. collection of documents related to Information Science. The documents were syntactically analyzed to generate highly associated document characteristics which were used to index the documents. The required data files were also described and algorithms for assigning disk addresses to text records were proposed. After the preliminaries of text processing were given, the actual document retrieval operations were defined and the various retrieval options including browsability were outlined.

At this point, I would like to make a few brief comments regarding the feasibility of such a document retrieval system. In terms of hardware, the system would require at least an on-line processing capability with remote terminal access. A time-shared computer system such as an IBM-360/67 or a GE-645, or in fact, any comparable processing system would suffice. A main core memory of no less than 8K words would be necessary for operating programs and for data space. As secondary storage, a large capacity disk storage unit would best satisfy the system's needs. For example, if we consider an IBM-2302/4 disk storage unit, which has a data capacity of 224,280,000 bytes (or alphanumeric characters), we could get some idea of how much space the retrieval system would require.

File	Millions Bytes	No. of Disks
Document Text File	130	35
Central Topic File	25	6
Serial File	2	.5
Inverted File	.1	.025
Term Table	.25	.05
System Programs	.5	.1

The estimates for the numbers of bytes in each file is based on an assumed average of six characters per English word. For example, if it is assumed that an average document in the C.I.S. file contains 10,000 words and there are 2,100 documents, then the total number of

characters (or bytes) in this document collection would be about 126 million.

Although it is apparent that the hardware exists and is available, it by no means implies that such a retrieval system could be implemented economically on such apparatus. However, it does seem reasonable to expect that the average cost per query would not be prohibitively high since the amount of processing time for each request would be low as a result of the very fast processors in time-shared systems.

References

[1] Hillman, D. J., "Characterization and Connectivity," Report No. 1, Document Retrieval Theory, Relevance, and the Methodology of Evaluation, National Science Foundation, Grant No. GN-451, 1966, p. 7, Lehigh University, Center for the Information Sciences.

[2] Hillman, D. J., Reed, D. M., "Microcategorization for Text Processing," Report No. 3, "Canonical Decomposition," Report No. 4, Document Retrieval Theory, Relevance, and the Methodology of Evaluation, GN-451, 1966, Lehigh University, Center for the Information Sciences.

[3] Ibid.

[4] Hillman, D. J., "The Measurement of Simplicity," Philosophy and Phenomenological Research, Vol. XIX, No. 4, 1959, pp. 429-446.

[5] Goodman, N., "About," Mind, Vol. LXX, No. 277, Jan. 1961, pp. 1-24.

[6] Mooers, Calvin N., "A Mathematical Theory of Language Symbols in Retrieval," Proceedings of the International Conference on Scientific Information, Washington, D. C., November 16-21, 1958, pp. 1327-1364.

[7] See References 1 and 2.

[8] Fairthorne, R. A., Towards Information Retrieval, Butterworth's, London, 1961, p. 127.

[9] Luhn, H. P., "Keyword-In-Context Index for Technical Literature (KWIC Index)," American Documentation, 11, 1960, pp. 228-295.

[10] See Reference 2.

[11] McCarthy, J., Abrahams, P. W., Edwards, D. J., Hart, T. P., Levin, M. I., "LISP 1.5 Programmers Manual," M.I.T. Computation Center, Cambridge, Mass., 1962.

[12] Farber, D. J., Griswold, R. E., Polonsky, I. P. "SNOBOL, A String Manipulation Language," J. ACM., 11, 1, Jan. 1964, pp. 21-30.

APPENDIX I

Dictionary Program and Listings

Input to the dictionary program is presently by cards. Periods are used only to mark end of sentences and are punched as separate words. Commas are also punched as separate words.

The program first compares each word of text with the functor word dictionary listed in Table A below. If a match is found, the text word is assigned the category listed for it in the dictionary. If no match is found, the program then compares the text word with the two suffix dictionaries.

For a suffix search, all final s's are deleted from the text word. This was done to shorten the suffix dictionaries and to facilitate programming. In the first suffix search the last two letters of the text word (minus s's) are compared with the two letter suffixes listed in Table B below. If no match is found, then the last three letters of the text word are compared with the three letter suffixes listed in Table C below. If a match is found in either dictionary, the corresponding category is assigned to the text word. If no match is found in any of the dictionaries, the word is assigned the type 'U' for unknown.

The text words followed by their categories form the output on tape. A period is assigned the category '.QQ', which functions as an end of sentence tag for the other programs.

Table A

Functor Word Dictionary (in machine alphabetic order)

Word	Category	Word	Category
About	P	Could	AV
Above	P	Didn't	AV
Across	P	Did	AV
Adequate	A	Doesn't	AV
After	P	Does	AV
Again	B	Don't	AV
Against	P	Do	AV
All	A	Each	A
Along	P	Either	C
Also	B	Else	N
Always	B	Everybody	N
Among	P	Everyone	N
Am	AV	Everything	N
And	C1	Except	P
An	ART	Fail	V
Anybody	N	Fails	V
Anyone	N	Fewer	A
Any	A	Fewest	A
Anything	N	Few	A
Apart	P	First	B
Apply	V	For	P
Aren't	AV	From	P
Are	AV	Hadn't	AV
Around	P	Had	AV
A	ART	Hasn't	AV
As	C3	Has	AV
At	P	Haven't	AV
Away	B	Have	AV
Back	P	Having	AV
Because	C	Hence	B
Been	AV	Her	N
Before	P	Here	B
Behind	P	Hers	A
Being	AV	Herself	N
Below	P	He	N
Be	AV	Him	N
Beside	P	Himself	N
Between	P	His	A
Beyond	P	However	B
Both	C2	How	B
But	C	If	C
By	P	Including	C1
Can't	AV	In	P
Cannot	AV	Inside	P
Can	AV	Into	P
Couldn't	AV	I	N

Table A (Cont'd.)

Word	Category	Word	Category
Isn't	AV	Sometimes	B
Is	AV	Somewhat	A
It	N	Still	B
Its	A	Such	A
Itself	N	Than	C2
.	QQQ	That	T
Many	A	Their	A
May	AV	Them	N
Mention	V	Then	C
Me	N	Therefore	B
Might	AV	There	N
More	A	The	ART
Mustn't	AV	These	N
Must	AV	They	N
My	A	This	N
Myself	N	Those	N
Near	P	Too	B
Need	N	To	P
Needs	N	Toward	P
Nobody	N	Under	P
None	N	Until	K
Nor	C	Upon	P
No	A	Up	P
Not	B	Us	N
Now	B	Very	B
Of	P	Via	P
Only	B	Wasn't	AV
On	P	Was	AV
Or	C1	Weren't	AV
Other	A	Were	AV
Others	N	We	N
Oughtn't	AV	When	C
Ought	AV	Where	C
Our	A	Which	K
Ours	A	While	C
Out	B	Whoever	N
Outside	P	Whom	K
Over	P	Who	K
Own	A	Whose	K
Per	P	Why	B
Rate	N	Will	AV
Rates	N	Within	P
Rather	C	Without	P
She	N	With	P
Shouldn't	AV	Wouldn't	AV
Should	AV	Would	AV
Since	C	Your	A
Somebody	N	Yours	A
Someone	N	Yourself	N
Some	N	You	N
So	B	,	C1
Something	N		

Table B

Two Letter Suffix Dictionary

Suffix	Category
CY	N
ED	VD
ER	N
IC	A
LY	B
UM	N
US	N

Table C

Three Letter Suffix Dictionary

Suffix	Category	Suffix	Category
AGE	N	ION	N
AIN	V	ISE	B
ANT	N	ISH	A
APH	N	ISM	N
ARD	A	IST	N
ARY	A	ITE	N
ATE	V	ITY	N
BLE	A	IVE	A
CIE	N	IZE	V
CUR	V	LAR	A
DOM	N	LTE	N
ECT	V	LOG	N
EDE	V	MIT	V
EED	V	NAR	A
ENT	A	NCE	N
ERY	N	OGY	N
EST	A	OID	N
ETH	N	OLY	N
ETY	N	ORY	A
EVE	V	OSE	V
FIE	V	OUS	A
FUI	A	PEL	V
GIE	N	RAM	N
GUE	N	RIE	N
HER	N	SAL	N
HIP	N	SCE	V
IAN	N	SHE	V
IAR	A	THE	V
IBE	V	TIE	N
IER	N	TOR	N
IFY	V	TRY	N
ILE	N	UCE	V
INE	N	UDE	N
ING	VP	URE	N

APPENDIX II

The following output is an example of the form of output produced by the syntactic analyzer used in text processing.

```
DCN  [illegible]
DXCXMX  ONE
INPUT   NP   PROPOSITIONAL   A   LOGIC   N   NP   IS   V   PRESENTED   V   PAD   HERE
A   PAD   AS   C3   NP   A   A   SYNTACTICAL   A   SYSTEM   N   NP   DESIGNED   VD
VI   TO   DD   PRODUCE   DD   VI   EXACTLY   B   NP   THOSE   A   FORMULAS   N   NP
QQQ
PROPOSITIONAL   A   LOGIC   N
A   A   SYNTACTICAL   A   SYSTEM   N
THOSE   A   FORMULAS   N
ENDFST
INPUT   NP   THOSE   A   FORMULAS   N   NP   REPRESENT   V   NP   VALID   A   PRINCIPLES
N   NP   ,   C1   NP   VALID   N   NP   VI   TO   DD   BE   DD   UNDERSTOOD   DD
VI   INTUITIVELY   B   BUT   C   REDUCED   VD   Y   TO   P   W   THE   A   STANDARD
A   TWO-VALUED   A   TRUTH-TABLES   N   W   Y   Y   FOR   P   W   ***   N   W
,   CP   W   *V*   N   W   ,   CP   W   ***   N   W   ,   CP   W   *#*   N
W   Y   QQQ
THOSE   A   FORMULAS   N
VALID   A   PRINCIPLES   N
VALID   N
W   THE   A   STANDARD   A   TWO-VALUED   A   TRUTH-TABLES   N   W   FOR   P   ***
N   W   ,   CP   W   *V*   N   W   ,   CP   W   ***   N   W   ,   CP   W
*#*   N
ENDFST
INPUT   NP   THE   A   LATTER   A   REDUCTION   N   NP   IS   V   SOMEWHAT   B   CONCEALED
VD   Y   BY   P   W   THE   A   FACT   N   W   Y   THAT   T   NP   TRUTH-TABLES
N   NP   ARE   V   NOT   B   TREATED   VD   UNTIL   C   INTRODUCED   VD   QQQ
THE   A   LATTER   A   REDUCTION   N
W   THE   A   FACT   N   W
TRUTH-TABLES   N
ENDFST
INPUT   NP   THIS   N   NP   IS   V   NP   A   A   TEST   A   CARD   N   NP
```

BEST AVAILABLE COPY

```
Y    TO    P    W    TESR    N    W    Y    NP    THE    N    NP    NP    THE    A    CONTINUATION
N    NP    UF    P    Y    CF    P    W    DATA    N    W    Y    Y    FROM    P    W
THREFRO    A    ONE    A    VCARD    N    W    Y    Y    TO    P    W    ANOTHER    N    W
Y    QQQ
THIS    N
A    A    TEST    A    CARD    N    TO    P    TESR    N
THE    N
THE    A    CONTINUATION    N
W    DATA    N    A    FROM    P    THREFRO    A    ONE    A    VCARD    N    TO    P    ANOTHER
N
ENDFST
DXCXMX    TWO
INPUT    NP    THIS    A    MANUAL    N    NP    IS    V    DESIGNED    V    PRIMARILY    B    Y
FOR    P    W    PHILOSOPHY    A    STUDENTS    N    W    Y    QQQ
THIS    A    MANUAL    N
W    PHILOSOPHY    A    STUDENTS    N    W
ENDFST
INPUT    NP    PHILOSOPHY    A    STUDENTS    N    NP    HAVE    V    NP    NO    A    PREVIOUS
A    KNOWLEDGE    N    NP    Y    IN    P    W    MATHEMATICS    N    W    Y    QQQ
PHILOSOPHY    A    STUDENTS    N
NO    A    PREVIOUS    A    KNOWLEDGE    N    IN    P    MATHEMATICS    N
ENDFST
INPUT    N    FOR    P    W    SUCH    A    STUDENTS    N    W    Y    ,    C1    NP    THIS
A    BOOK    N    NP    SHOULD    V    PROVE    V    PAD    VERY    B    SATISFACTORY    A    PAD
QQQ
THIS    A    BOOK    N
W    SUCH    A    STUDENTS    N    W
ENDFST
INPUT    NP    ITS    A    CONTENT    N    NP    HAS    V    BEEN    V    PAD    CAREFULLY    B
CHOSEN    A    PAD    B    B    AS    C3    VI    TO    DD    GIVE    DD    VI    NP    AN    A
INTRODUCTION    N    NP    Y    TO    P    W    THE    A    FUNDAMENTALS    N    W    Y    Y
OF    P    W    MODERN    A    ELEMENTARY    A    LOGIC    N    W    Y    QQQ
ITS    A    CONTENT    N
```

BEST AVAILABLE COPY

```
AN    A    INTRODUCTION   N    TO    P    THE   A    FUNDAMENTALS   N    OF    P    MODERN   A
ELEMENTARY    A    LOGIC    N
ENDFST
INPUT    NP    A    A    LIST    N    NP    Y    OF    P    W    EXERCISES   N    W    Y
AND   C1    NP    A    A    SELECTED    A    BIBLIOGRAPHY   N    NP    ARE   V    INCLUDED    V
QQQ
A    A    LIST    N    OF    P    EXERCISES   N
ENDFST
A    A    SELECTED    A    BIBLIOGRAPHY   N
ENDFST
INPUT    N    IN    P    W    THE   A    FIRST    A    CHAPTER    N    W    Y    ,    C1
NP    SENTENTIAL    A    LOGIC    N    NP    IS    V    INTRODUCED    V    AS    C3    NP    A
A    SYSTEM    N    NP    Y    OF    P    INTERPRETED    PVD   W    SIGNS    N    W    Y    QQQ
SENTENTIAL    A    LOGIC    N
A    A    SYSTEM    N    OF    P    SIGNS    N    IN    P    THE   A    FIRST    A    CHAPTER
N
ENDFST
INPUT    NP    THE   A    MAIN    A    PURPOSE    N    NP    IS    V    NP    THE   A    PRESENTATION
N    NP    Y    OF    P    W    THE   A    CONCEPT    N    W    Y    NP    THIS    A    PROGRAM
N    NP    IS    V    PAD   SUCCUSFUL   A    PAD   QQQ
THE   A    MAIN    A    PURPOSE    N
THE   A    PRESENTATION   N    OF    P    THE   A    CONCEPT    N
THIS    A    PROGRAM    N
ENDFST
INPUT    NP    THIS    A    CARD    N    NP    IS    V    Y    IN    P    W    THIS    A
DECK    N    W    AND   CP    W    PROGRAM    N    W    Y    AS    C3    NP    A    A    TEST
N    NP    QQQ
THIS    A    CARD    N
W    THIS    A    DECK    N    W    AND   CP    W    PROGRAM    N    W
A    A    TEST    N
ENDFST
DEC 2[illegible] 6[illegible]    13 30.3
```

BEST AVAILABLE COPY

APPENDIX III

The following matrices and their transformations are examples of the connectivity assignment procedures that are used in the text analysis operations.

To illustrate the document characterization process let $C = S,U,V,W,X,Y,Z$ be the document collection and suppose that $P = a,b,c,d,e,f,g,h,i,j,k,l,m,n$ is the repertory of document characteristics for the document collection C. Suppose also that the documents in C have the following relational structure:

$$S: \{R_1(c),\ R_2(d),\ R_3(g,j)\}$$

$$U: \{R_4(a,h),\ R_5(j)\}$$

$$V: \{R_6(h,n),\ R_7(h,l),\ R_8(n)\}$$

$$W: \{R_9(a),\ R_{10}(k)\}$$

$$X: \{R_{11}(i,l,n),\ R_{12}(l,n)\}$$

$$Y: \{R_{13}(b),\ R_{14}(f)\}$$

$$Z: \{R_{15}(e),\ R_{16}(e,m)\}$$

The affiliation matrix, the incidence matrix, and the submatrices resulting from the partitioning of the incidence matrix along with their respective unique probability vectors are shown on the following three pages.

CHARACTERISTIC X DOCUMENT MATRIX

2	2	0	1	0	2	0
0	0	0	0	0	1	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	3
0	0	0	0	0	1	0
2	0	0	0	0	0	0
0	2	4	0	0	0	0
0	0	0	0	3	0	0
2	1	0	0	0	0	0
0	0	0	1	0	0	0
0	0	2	0	5	0	0
0	0	0	0	0	0	2
0	0	3	0	5	0	0

CORRESPONDING INCIDENCE MATRIX

5	0	0	0	0	0	0	4	0	2	1	0	0	0
0	1	0	0	0	1	0	0	0	0	0	0	0	0
0	0	1	1	0	0	2	0	0	2	0	0	0	0
0	0	1	1	0	0	2	0	0	2	0	0	0	0
0	0	0	0	9	0	0	0	0	0	0	0	6	0
0	1	0	0	0	1	0	0	0	0	0	0	0	0
0	0	2	2	0	0	4	0	0	4	0	0	0	0
4	0	0	0	0	0	0	20	0	2	0	8	0	12
0	0	0	0	0	0	0	0	9	0	0	15	0	15
2	0	2	2	0	0	4	2	0	5	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	8	15	0	0	29	0	31
0	0	0	0	6	0	0	0	0	0	0	0	4	0
0	0	0	0	0	0	0	12	15	0	0	31	0	34

SUBMATRICES OF THE PARTITIONED INCIDENCE MATRIX

A SUBMATRIX

5	0	0	0	4	0	2	1	0	0
0	1	1	2	0	0	2	0	0	0
0	1	1	2	0	0	2	0	0	0
0	2	2	4	0	0	4	0	0	0
4	0	0	0	20	0	2	0	8	12
0	0	0	0	0	9	0	0	15	15
2	2	2	4	2	0	5	0	0	0
1	0	0	0	0	0	0	1	0	0
0	0	0	0	8	15	0	0	29	31
0	0	0	0	12	15	0	0	31	34

ROW	ROW SUM
1	12
3	6
4	6
7	12
8	46
9	39
10	17
11	2
12	83
14	92

THE UNIQUE PROBABILITY VECTOR

3.8095238-02
1.9047619-02
1.9047619-02
3.8095238-02
1.4603175-01
1.2380952-01
5.3968254-02
6.3492063-03
2.6349206-01
2.9206349-01

A SUBMATRIX
1 1
1 1

ROW	ROW SUM
2	2
6	2

THE UNIQUE PROBABILITY VECTOR
5.0000000-01
5.0000000-01

A SUBMATRIX
9 6
6 4

ROW	ROW SUM
5	15
13	10

THE UNIQUE PROBABILITY VECTOR
6.0000000-01
4.0000000-01

PHRASE INDEXING

by

David M. Reed

Abstract

In Part I, a manual indexing system, using phrases rather than uniterms or descriptors, is developed and evaluated in terms of certain assumptions about user oriented systems. Part II deals with retrieval operations for manually and automatically phrase indexed systems.

This research was supported by the National Science Foundation under Grants No. GN-451 and No. GE-2569

Preceding Page Blank

Table of Contents

Page

Part I.....3-1

Introduction.....3-1

Part II.....3-5

A. Preliminaries.....3-5

B. Generation of Files for Retrieval Purposes.....3-5

C. Outline of Retrieval Operations.....3-8

D. Flowcharts.....3-11

PART I. Introduction

The major guiding principle in the conception and design of this IR system is to produce a user-oriented system. It is assumed that in a user-oriented system that the user should not be expected to know about the technicalities of the system, that he should not be forced to express his request in very restricted and structured vocabulary and form, and that his interface with the system should be in normally meaningful natural language.

To satisfy these assumptions indexing of documents in the system is done by assigning short descriptive phrases to documents. Phrases rather than uniterms or descriptors are used because uniterms and descriptors by themselves have little definite meaning to someone not well acquainted with the particular system at hand. It is felt that phrases provide contextual meaning for terms embedded in them and that this contextual meaning of terms will provide for meaningful interface between the system and the user.

The generation of these descriptive phrases by the indexer is governed by a semi-controlled vocabulary, possibly a few syntactic constraints, and an acceptability check by the system. The indexer is given a basic word vocabulary consisting of the general vocabulary of the field of the documents. The indexer has the liberty to augment the basic word vocabulary with other words for the purposes of additional qualification, e.g., in the phrase "Goodman's concept of relevance," "concept" and "relevance" would be basic vocabulary and "Goodman" augmented vocabulary. The system is so designed that augmented vocabulary will not have adverse effects on the system's performance.

There are no theoretical restrictions on the number of phrases that can be assigned to a document. Because of this and other factors discussed below, this system has parallels with Hillman's document characterization system [1]. Obviously the more phrases assigned per document the more detailed the indexing is.

Syntactic constraints on indexing phrases and the indexing acceptability check by the system are discussed below.

The indexing phrases are interpreted by the system as graph connected word strings. Each non-trivial word in an indexing phrase is taken to be a graph node. A phrase is represented as a connected path in linear order between the words in the phrase. The graph is then represented in terms of a matrix.

The structuring of indexing phrases in terms of graph theoretic concepts was chosen on the basis of the following considerations:

1. It is desirable to structure the indexing phrases in a form which captures the maximal structure of the phrases themselves because it is assumed to be desirable not to impose any more structure upon them than they actually possess. It is felt that by structuring the phrases in the most general form they possess the system does not make a priori commitments to any particular theoretical structure the phrases might be thought to have. Since the system does not incorporate at the grass roots level any structural model except a most general weak one, one is left free in the retrieval processes to impose upon the phrases a wide choice of structural models. Retrieval processes in the system thus can manipulate the phrases as is pragmatically useful because of the lack of any strong structuring in the storage of the indexing phrases. Since the phrases are interpreted in a way that captures only their most general structure, the structural assumptions of any particular retrieval process remain explicitly clear and unconfused with the structure of the indexing phrases.

2. It is believed that a graph interpretation of the indexing phrases provides all the structure necessary for a retrieval system to satisfy the assumptions made at the beginning of this paper about user-oriented systems.

3. This technique of structuring the indexing phrases superimposes a word-word matrix on a phrase-phrase matrix similar to the term-term matrix generated by Hillman's technique [2]. Hillman's process generates a phrase-phrase matrix based on phrase co-occurrence and a weighting factor. All phrases used to characterize a particular document with manual indexing as well as in Hillman's system of document characterization are bound in the same genus. All the phrases of other documents which have at least one phrase in common with this one document are also bound in the same genus. Although this process groups together conceptually related documents by grouping together their characteristics, it does not provide for user-oriented retrieval techniques since there is little likelihood that a user will give in his request an exact phrase which was used to characterize a document.

However a word-word matrix superimposed on the phrase-phrase matrix allows retrieval techniques which lead the user heuristically to the manually produced phrases or the source oriented phrases of Hillman's system which are appropriate to the user's request statements.

Part II of this paper develops this system in detail. It deals explicitly with retrieval operation processes for a document collection characterized by Hillman's methods. However in this part of the paper I have dealt with a retrieval system in which document characterizations are manually produced. Given documents manually indexed with descriptive phrases a term-term (i.e. phrase-phrase) matrix and a term (phrase)-document matrix can be generated and these matrixes form the input to that part of the system described in Part II. I believe that everything in Part II is applicable to manual indexing systems of the type discussed above as well as to Hillman's system. In addition to the considerations in Part II there are two other facets of such manual indexing systems which are particularly significant for system design and evaluation.

Since in manual systems the syntax of the indexing descriptive phrases is more controllable than in automatic systems, the utility of restrictive or special meaning phrase syntax can be more fully examined. (See Part II)

In a manual indexing system the retrieval operation procedures can provide an acceptability check on the indexing. This acceptability check can evaluate the effect of new phrases on the structure of the term-term matrix and the word-word matrix. If a set of new indexing phrases links previous distinct genera (partitions) of the term-term matrix, the indexer, using the browsing operations described in Part II, can determine if the new documents as indexed really relate the two different conceptual areas of the distinct genera and/or if it is desirable from a subject matter point of view to form the two genera into one genus. This check thus allows manual structuring of the conceptual areas defined by genera.

Operations with the word-word matrix can present the indexer with a resumé of the contextual meaning previously given to the indexing vocabulary. The indexer thus can determine if a set of new indexing

phrases employs vocabulary consistent with past use of vocabulary. If it is found desirable to modify the meaning of vocabulary words, the system can readily present the past usage of selected words for appropriate decisions. Such modifications in vocabulary usage will require of course the regeneration of the files used for retrieval operations but no change in retrieval operations. Additions to the indexing vocabulary can be checked for subject matter consistency by examination of the genera in which the system places the phrases containing the new vocabulary.

PART II.

A. Preliminaries

The text processing and matrix operations described above yield a document-term matrix and a partitioned term-term matrix which together characterize a document collection. To utilize this characterization fully for document retrieval it is necessary to develop retrieval techniques which make optimal use of all of the information about the system's documents contained in these matrices. Since there is little likelihood that a user will employ exactly the same terms in requests as the document characterization procedure or the indexer selects to characterize documents, the problem of interface between user and system is of central importance for making optimal use of the matrices. To make minimal demands upon the user a word-word matrix, for each non-trivial word occurring in document characterizations, is generated and is used in retrieval operations to heuristically lead the user from the terminology of his request statements to the source derived or indexer supplied terminology of the document characterization.

It is believed that this technique as discussed in the next two sections will successfully allow useful and meaningful interface between the user and the information the system contains about its documents and will allow the user to make full use of document characterization associations in a intuitively intelligent manner.

An attempt has been made in the following discussion to delineate the generation and structure of data files and the retrieval operation procedures in a fashion which is easily programmed. Flowcharts of proposed programs for these are below. The system configurations needed are indicated on the flowcharts.

B. Generation of Files for Retrieval Purposes

Input to these programs consists of the partitioned term-term matrices and the document-term matrix described above. The terms of the matrices are source derived phrases. Each phrase is assigned a unique code number which is called a "String Number." A double entry dictionary is constructed containing string numbers and their correlate phrases. Following each phrase is its entry from the document-term

matrix which contains the documents the phrase characterizes and its width in each of these documents. This dictionary is called the "String Number-Phrase Dictionary."

STRING NUMBER	SOURCE DERIVED PHRASE	DOCUMENT NUMBER	WIDTH OF PHRASE IN DOC.	DOCUMENT NUMBER	WIDTH OF PHRASE IN DOC.

Figure 1. String Number-Phrase Dictionary

Each genus in the term-term matrices is assigned a unique number and File A is generated. File A consists of the String Number of a phrase and the Genus Number of the genus in which it occurs.

STRING NUMBER	GENUS NUMBER

Figure 2. File A

File A and the String Number-Phrase Dictionary are merged to generate File B. The String Number-Phrase Dictionary is saved for further use. File A is discarded.

SOURCE DERIVED PHRASE	GENUS NUMBER	STRING NUMBER

Figure 3. File B

From File B is generated File C which consists of, for each non-trivial word in each Source Derived Phrase in File B, that word followed by its phrase's Genus Number and String Number. File B is discarded.

WORD	GENUS NUMBER	STRING NUMBER

Figure 4. File C

File C is sorted alphabetically on the words and file entities with identical words are combined to produce the Word Profile File as in Figure 5. File C is discarded.

WORD	GENUS NO.	STRING NUMBERS	GENUS NO.	STRING NUMBERS

Figure 5. Word Profile File

The input partitioned term-term matrices are rewritten using the String Number-Source Phrase Dictionary to produce the String Association Matrix File. The input matrices are now discarded.

GENUS NO.	STRING NUMBER	ASSOCIATED STRING NUMBER	ASSOC. VALUE	ASSOCIATED STRING NUMBER	ASSOC. VALUE

Figure 6. String Association Matrix File

This file is ordered by Genus Number and String Number.

A subsidiary file needed for document identification is a Document Number - Bibliographic Data Dictionary which contains the title, author, source, etc. identification of documents. (See Figure 7)

Figure 7 contains the complete set of files used for retrieval purposes.

STRING NUMBER	SOURCE DERIVED PHRASE	DOCUMENT NUMBER	WIDTH OF PHRASE IN DOC.	DOCUMENT NUMBER	WIDTH OF PHRASE IN DOC.

String Number-Phrase Dictionary

WORD	GENUS NUMBER	STRING NUMBERS	GENUS NUMBER	STRING NUMBERS

Word Profile File

GENUS NUMBER	STRING NUMBER	ASSOCIATED STRING NUMBER	ASSOCI-ATION VALUE	ASSOCIATED STRING NUMBER	ASSOCI-ATION VALUE

String Association Matrix File

DOCUMENT NUMBER	BIBLIOGRAPHICAL DATA

Document Number-Bibliographical Data Dictionary

Figure 7. Files Used For Retrieval

C. Outline of Retrieval Operations

The following retrieval operations are designed to be as user-oriented as possible. It is assumed that it is desirable in a user-oriented system not to force the user to express his initial request in a very restricted vocabulary and form. The user therefore submits his initial request in the form of a phrase describing what he wishes to find a document about. He may use any vocabulary he wishes. The system heuristically develops his initial request through interface with the user to obtain source derived phrases which the user indicates as elaborations and refinements of his initial request. The user can be lead by the system to browse in the general area of his request and to broaden or narrow his request as he wishes. The user may consult documents related to his request and by accepting and rejecting them modify

his description of what he is looking for.

Interface between the system and user is in natural language and the user is not required to know about the technicalities of the system. It is felt that these factors are highly desirable.

In the actual operation of the system there are many options possible for user-system interaction in the determination of the flow of the retrieval processes. Experimental evaluation of the several retrieval operation flow patterns must be made to determine the optimal patterns in terms of the user's satisfaction with retrieved documents and with the ease of interacting with the system. In the following outline of proposed retrieval operation flow such options will be noted.

When the user has entered his request in the form of a description, the request is scanned for words which occur in the Word Profile File. It is an open question if a syntactic analysis of the request will furnish any information which would improve the system's understanding of the request, e.g., Will restrictive clauses in the request usefully narrow the system's search? Will word order be important? Or will later user-system interaction resolve such problems more easily?

The profile records of words from the Word Profile File found in the request description are compared to determine if they all fall within one genus. If the request words are homogeneous then the operations described below are performed. If the words found in the request are found in different genera, the user is presented with a list of phrases from the different genera in which his words appear. The user is asked to make a selection of the phrase most pertinent to his request. If he is unsure or cannot decide, the phrases are presented again but with their most highly associated phrases from the String Association Matrix File. If the user is still undecided or wishes to explore the conceptual area of one choice, associated phrases of any phrase are presented until a choice of one is made or until a group of phrases from one genus is selected. The user may return to this point from further on in the operation if he feels that his original choices were incorrect or if he wishes to explore alternatives. It is believed that this reiterative process will operationally be of great value, allowing the user to browse from one genus to another as well as in a

genus and enabling the user to reformulate his original request in the terminology and phraseology of the system.

Evaluation of this browsing operation will concern how far from the original request it is practical to proceed and how meaningful and useful it would be to allow the user to rank presented phrases as pertinent to his wants instead of either accepting or rejecting them. Of over all interest will be comparisons of what the user thought originally he wanted and what the system leads him to formulate as his request, particularly when his reformulated request yields documents with which he is satisfied. It is believed that this browsing operation will transform whatever formalized "subject heading" nature initial requests have to either the point of view of the human indexer or the textual nature of the source-derived phrases, with a minimum of user discomfort and a maximum of informative user interface with the system.

It should be noted that the words in the phrases presented to the user are given contextual meaning by being embedded in phrases. This fact should insure that the user is presented by the system with meaningful expressions and meaningful choices.

The user arrives at the next stage of the retrieval operation having made a choice of a phrase or of phrases which he believes are the topic or topics he wishes to have a document about. The system now presents him with phrases which are directly associated with his reformulated request and the user has the option of further refining his request by adding additional phrases to his request and/or browsing a bit.

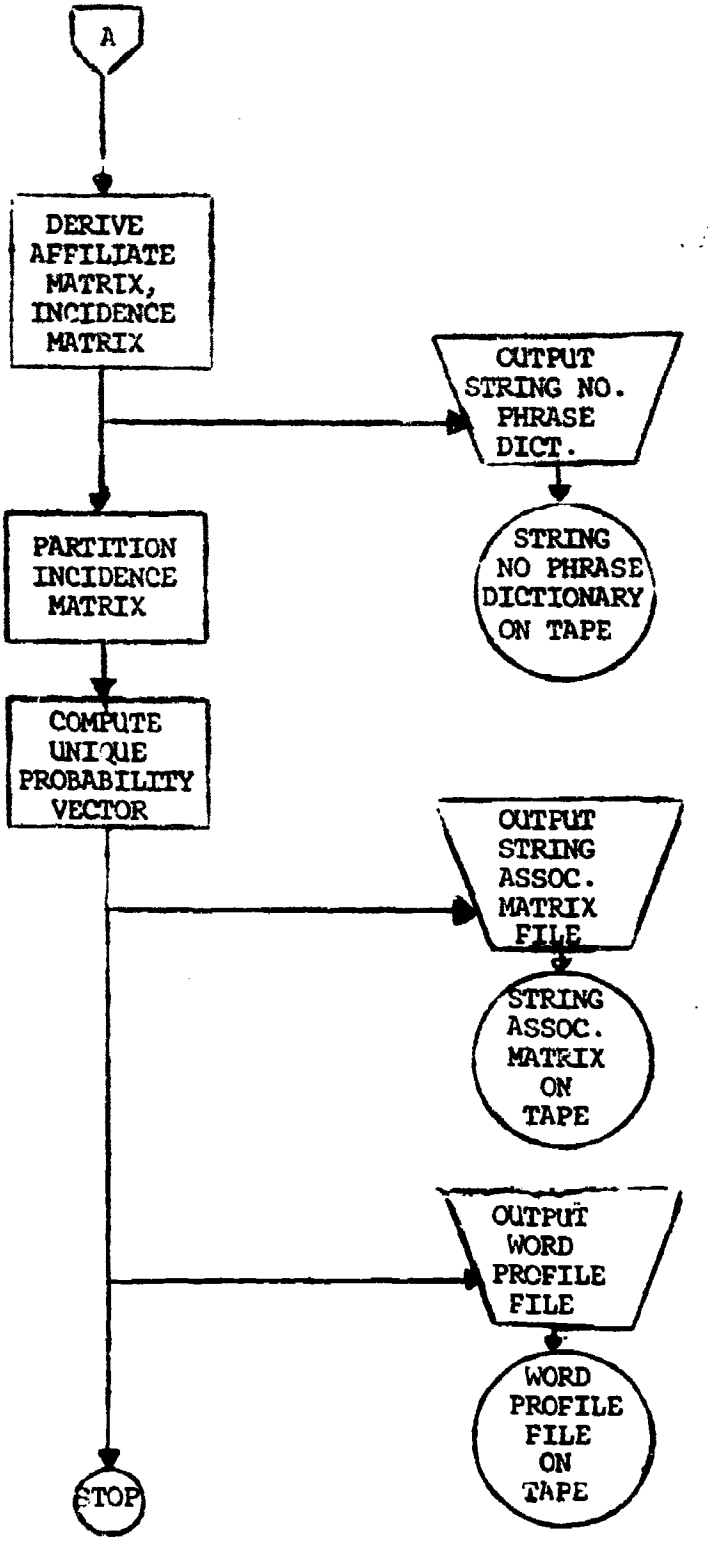
When the user has finally formulated his request, the system, using the String Number-Phrase Dictionary containing document numbers and widths, ranks pertinent documents numbers on the basis of maximal width values of the set of request phrases. The user is given the ranking of the documents and the necessary bibliographic data.

If the user wishes, he may ask for a display of selected passages from these documents and determine if he is satisfied. If he is not he may re-enter the retrieval operation, indicate those documents which are satisfactory and those which are not and browse in that portion of the

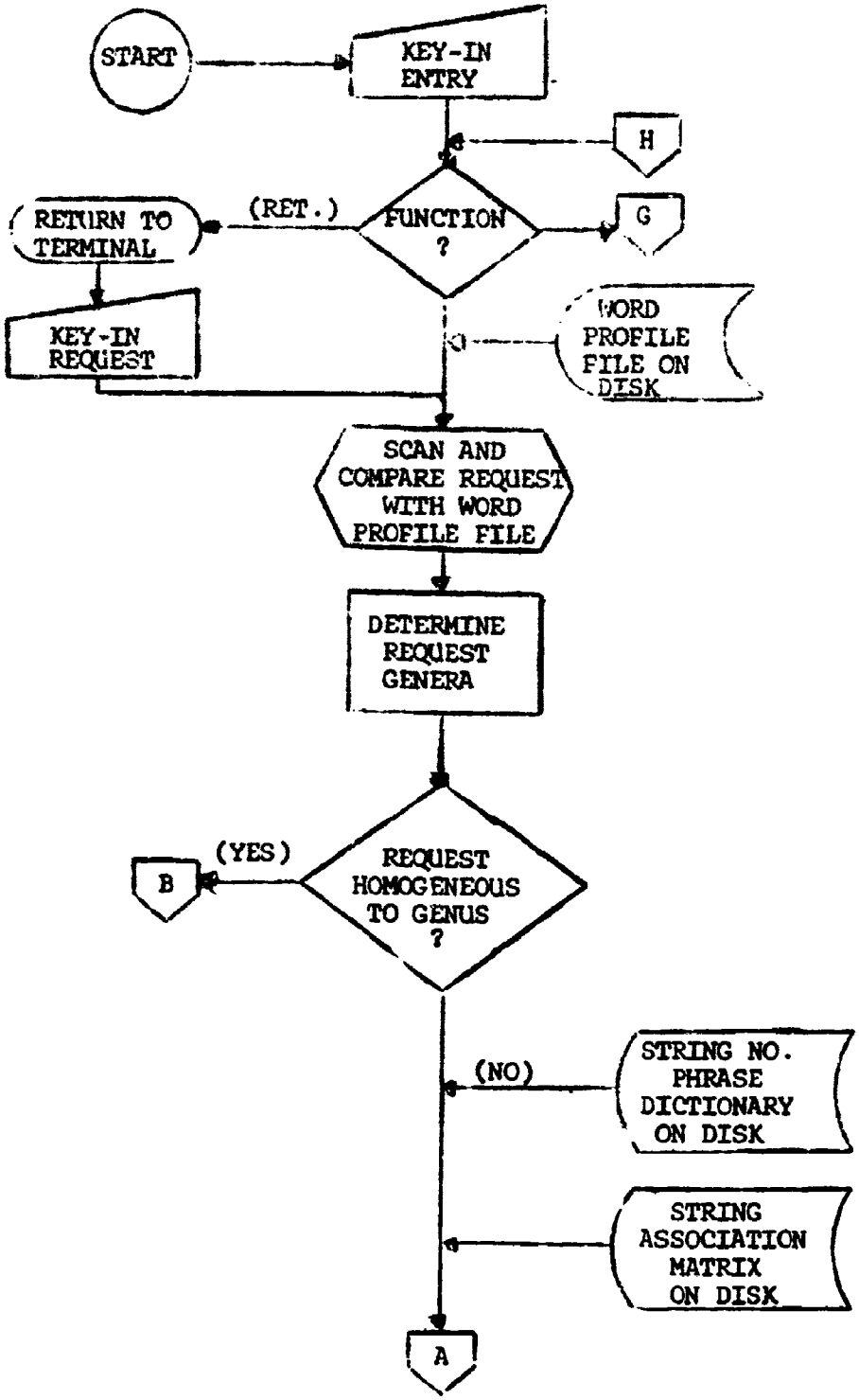
genus that contains the satisfactory documents. These operations are undertaken by the system by retaining the phrases which are associated with the satisfactory documents, inhibiting phrases associated with the rejected documents and returning to the browsing portion of the retrieval operation.

D. Flowcharts

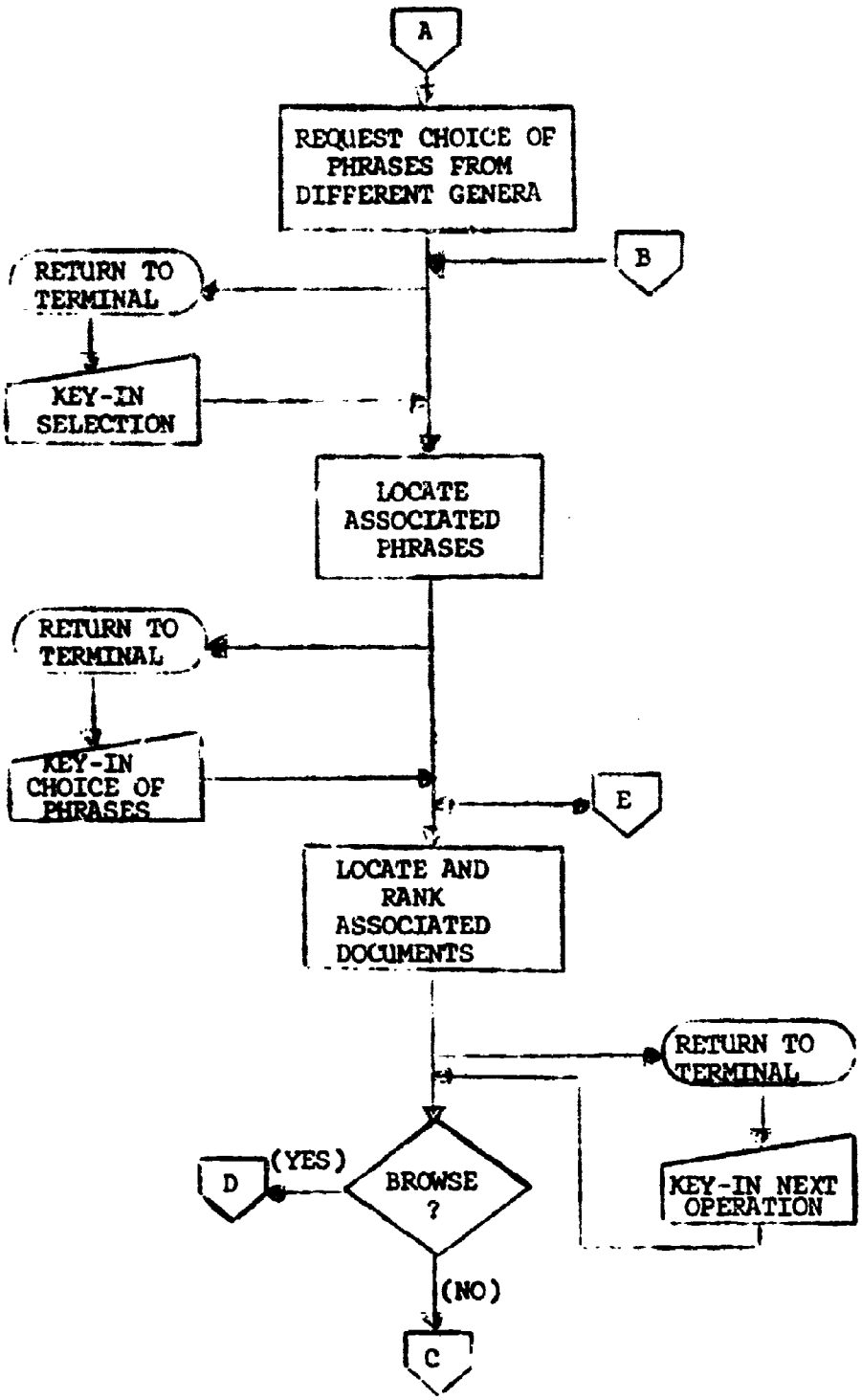
The following flowcharts summarize the generation of files used for retrieval and the flow of retrieval operations. I am indebted to Andrew Kasarda who drafted them for me.



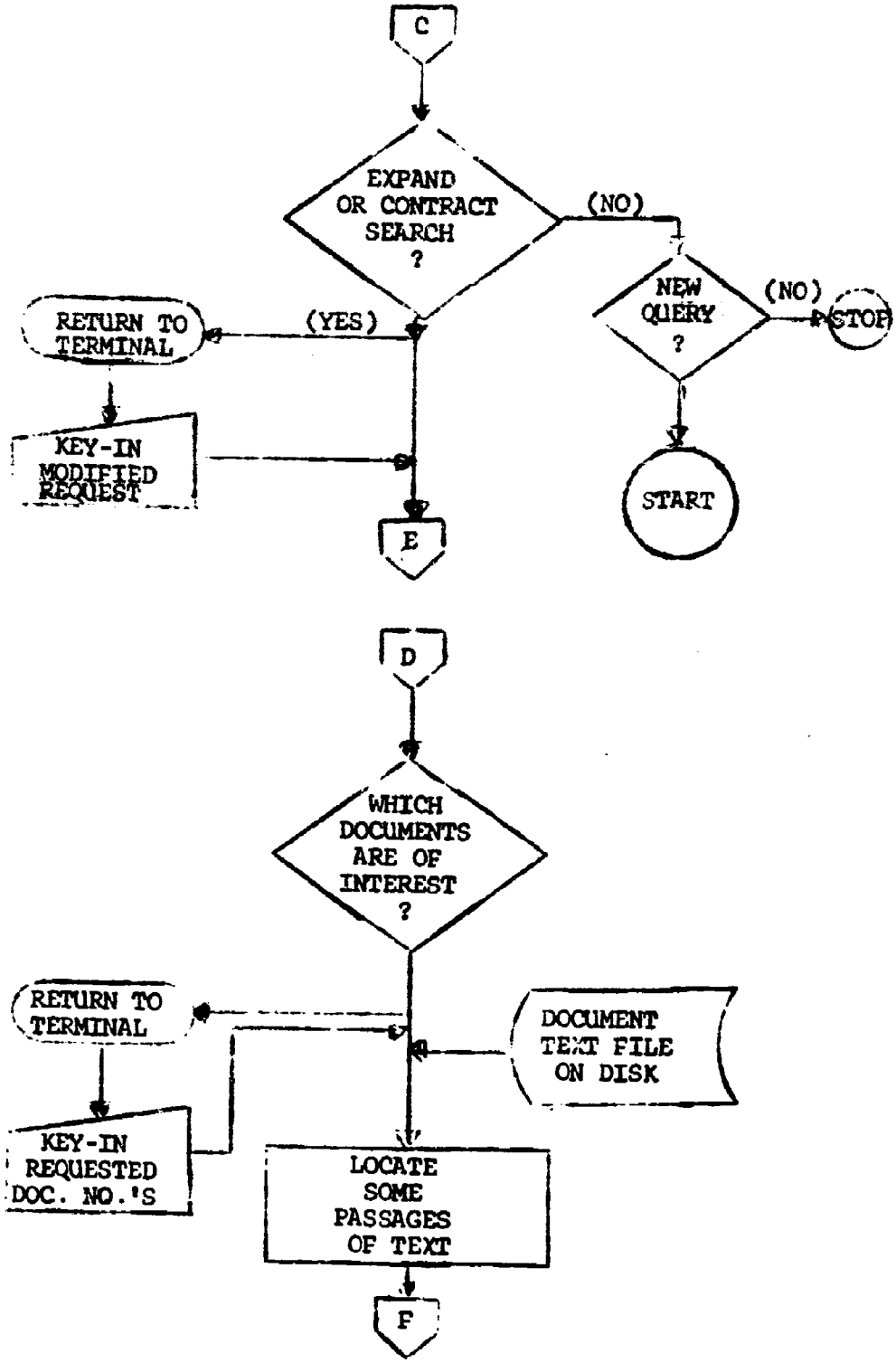
FILE PROCESSING SYSTEM



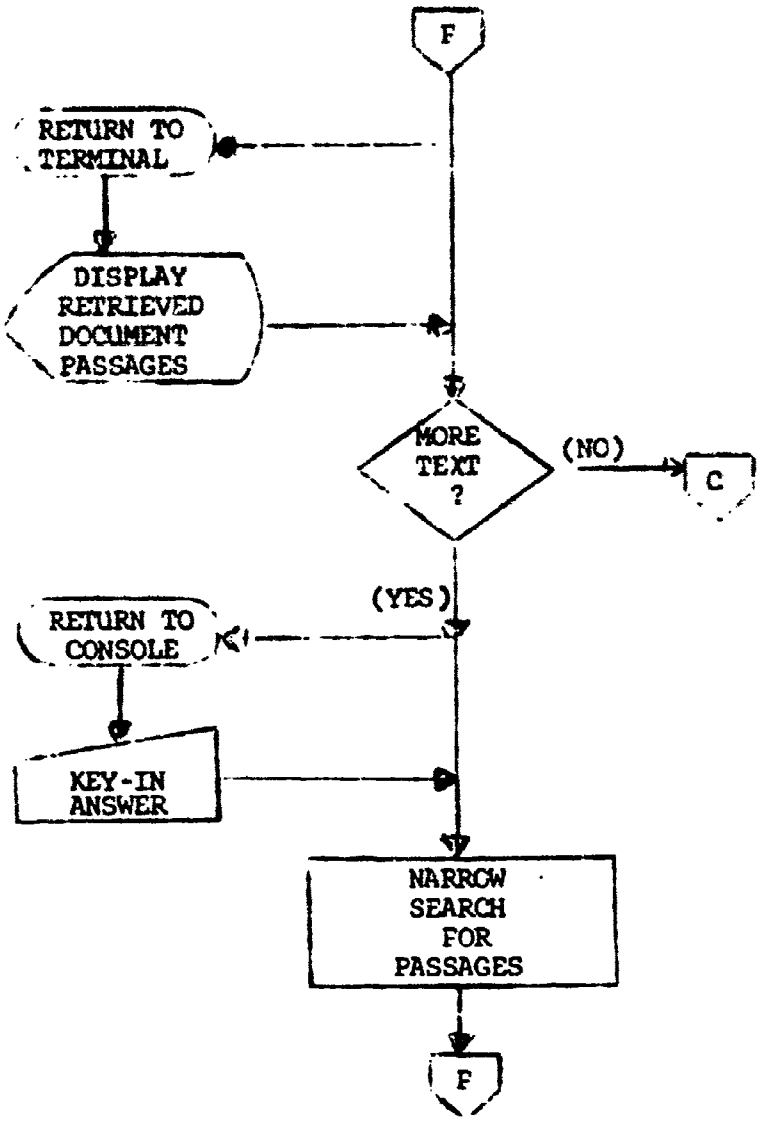
RETRIEVAL OPERATIONS - (PART-A)



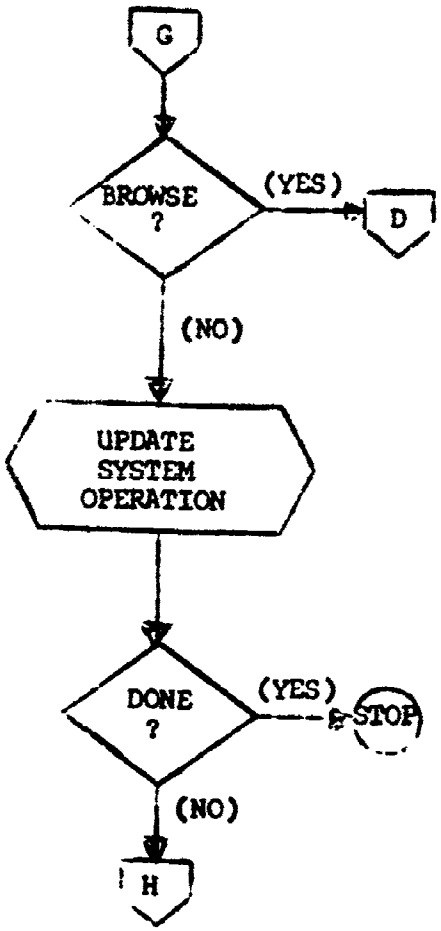
RETRIEVAL OPERATIONS - (PART-B)



RETRIEVAL OPERATIONS (PART-C)



RETRIEVAL OPERATIONS - (PART-D)



RETRIEVAL OPERATIONS - (PART-E)

References

[1] Donald J. Hillman, "Characterization and Connectivity," Document Retrieval Theory, Relevance, and the Methodology of Evaluation, NSF Grant No. GN-451, May 24, 1966.

[2] Ibid, pp. 18-38.